Updated: January 12, 2026

Winter Guerra

☎ | +1 (917) 435-7128
✉ | winterg@alum.mit.edu
⚡ | https://winter.industries/

Education

Massachusetts Institute of Technology 2019
Masters of Engineering in EE/CS (4.5/5.0)
Massachusetts Institute of Technology 2017
Bachelor of Science in EE/CS (4.1/5.0)

Work & Academic Experience

Millennium August 2024-Present

Team Lead: Applied AI

- Designed and developed a Deep Research multi-agent system using containerized open source models self-hosted on Kubernetes. Researches up to 400 sites per run in 18 minutes, achieving 1,600% lower cost than OpenAI Deep Research with comparable GAIA benchmark performance. Now used firm-wide by traders, quants, and division directors.
- Developed a deep-agent system with federated subagents using VLMs and browser automation (Playwright) to monitor hundreds of daily firm-wide support tickets from email and Slack. Provides L1 support and question answering, eliminating $1.6M+ in annual support expenditure.
- Architected federated, horizontally scalable agent-mesh systems for agent-to-agent interactions. Extensible architecture allows teams to add tools via ServiceNow forms or specialist agents. Now serves as design reference by infosec for secure multi-agent architectures.
- Architected and deployed a unified API layer for all internal and third-party LLMs, abstracting provider-specific APIs and authentication flows; eliminated redundant logic across teams and cut onboarding time from weeks to minutes. Now powers most high-volume generative AI applications including trading.
- Built a latency-and capacity-aware routing layer that dynamically selects optimal model endpoints across providers, reducing average time-to-first-token by 43% and increasing reliability for latency-sensitive applications used by traders and portfolio managers.
- Led development of a RAG-based multi-agent system, leveraging agentic code generation to autonomously run PostgreSQL queries across complex, internal datasets; enabled engineers and executives to obtain answers to ad-hoc questions without needing domain knowledge, reducing time-to-answer from hours to minutes.
- Architected onboarding and optimization patterns for self-hosted LLMs, reducing dependency on external providers and improving cost control and compliance posture for sensitive workloads.
- Spearheaded infrastructure and API design for long-running autonomous agents, including Kubernetes orchestration, caching, resumable execution, and seamless chat-based interaction—laying the technical foundation for future research and analysis agents.
- Drove the creation of a shared orchestration framework for multi-agent systems, establishing internal composability standards that accelerated agent development and enabled cross-team reuse of AI capabilities across the organization.
- Implemented core enterprise-grade AI security measures, including prompt injection defenses, sandboxed code execution, and scoped access controls—ensuring safe and compliant deployment of generative systems across the firm.

Reflex Robotics July 2024-August 2024

Consultant: AI, Robotics, & Computer Vision

- Architected a modular, microservice-oriented system design using high-speed in-memory IPC, replacing a monolithic C++ architecture; oversaw implementation by engineering team, now deployed in production.
- Developed a real-time depth estimation pipeline by temporally fusing deep learning-based disparity prediction with Semi-Global Matching to yield metrically scaled and temporally consistent depth maps.
- Engineered a VR rendering pipeline that compensates for stereo extrinsics, camera latency, and mismatch between human head motion and robot kinematics, eliminating operator motion sickness in teleoperation tasks.
- Designed and implemented a one-click sub-pixel accurate stereo calibration C++ tool using nonlinear optimization over an Extended Unified Camera Model (EUCM), enabling robust calibration of fisheye lens stereocameras with severe misalignment.
- Authored technical roadmap for visual odometry and multi-modal sensor fusion, enabling autonomous obstacle avoidance and reduced operator cognitive load; roadmap subsequently implemented by engineering team.
- Created a one-click calibration and rendering toolkit for rapid deployment, enabling automatic generation of dense depth maps and VR environment warping aligned with the user's ego-centric perspective.

Oishii — July 2021-March 2024

Head of AI, Computer Vision, & Robotics Team

- Spearheaded R&D and at-scale deployment of 3D robotic ML perception pipeline from end to end. Deployed on 48 6-DOF robots in semi-structured, highly-cluttered, harsh environments with 24/7 operation, processing over 1,200 RGB-D images per second.
- Invented novel cloud-native, high-throughput ML-based robotics deployment colocated with an on-premise high-availability Kubernetes cluster for increased accuracy, reliability, and scalability.
- Established AI, Computer Vision, and Robotics team as the company's first robotics/AI engineer. Built and led the team end-to-end: established technical viability through MVPs, designed recruiting and technical testing process, hired 6 engineers over 15 months.
- Designed and deployed novel deep learning-based method for identifying sub-millimeter-scale insects on gigapixel images. Beats PhD experts in recall percentage on unseen data, currently running in all production facilities nationwide at-scale.
- Architected 5 battle-tested CI/CD automation pipelines for central control of 1,800+ ML, robotics, and software docker containers across all facilities and IoT devices.
- Architected and deployed IoT computer-vision system for high-speed control and monitoring of bee activity in beehives, deployed in all production facilities across the USA on cost-effective resource-constrained hardware.
- Developed high-speed, automatic, single-pass eye-on-hand RGB-D camera calibration method for sub-minute, sub-millimeter accurate full robot calibration in the field by technicians.
- Co-designed robot-to-PLC-to-vision backend communication protocol using industrial Ethernet/IP.
- Debugged, tested, and modified FANUC on-robot code in production harvesting environments to improve robotic harvesting and automated calibration performance.
- Designed on-prem multi-facility physical networking architecture, class A IP addressing, subnets, fiber optic structured cabling, LACP links, failover, cluster-to-cluster VPNs, client VPNs, and programatically-assigned static routes using Ansible for handling over 48 Gbps of constant load.
- Standardized IoT platform software for rapid deployment of new embedded devices by technicians in a simple plug-and-play fashion. Simply insert a SD card into the IoT device, plug it into a PoE switch and the device will handle the rest!

SRI International — June 2020-July 2021

Computer Scientist II, Scene Understanding and Navigation Group

- Developed ML systems for external three-letter agency contracts; held security clearance.
- Architected parallel, GPU-accelerated ML Kubernetes pipeline for multi-modal attention-based image training, inference, and embedding search on 10 terrabyte dataset.
- Accelerated model training time from 3 days to less than 6 hours using GPU accelerated transform methods, cache-prediction, and deriving optimal mathematical calculation methods for internal model variables.
- Improved model recall by 600% through dynamic, GPU-accelerated intelligent sampling of strong positive and negative examples on every training rollout.
- Accelerated feed-forward image embedding inference and indexing from 4 days to 78 minutes.

MIT AeroAstro — 2017-2019

MEng Student in visual state estimation, planning, and simulation for UAVs in aggressive flight. Advised by Prof. Sertac Karaman

- Published novel, in-the-loop photorealistic virtual camera system for testing UAV visual state estimation and control in indoor agile flight using a high-speed monocular camera and IMU. Typical real world flight conditions exceeded $\geq 30.9mph$, $\geq 2.8G$, and $\sim 180\,\mathrm{Hz}$ for the onboard & virtual cameras.
- Co-published a paper on active perception methods for greatly increasing Visual-Inertial Odometry (VIO) accuracy and robustness in extremely challenging indoor environments using saliency maps of the environment.
- Lead development and design of a publically released Docker-based robotics simulator and automated scoring system used by over 500 competing teams in the 2019 AlphaPilot/Lockheed Martin AI Drone Racing Innovation Challenge ($1M grand prize).
- Published a novel fully annotated 4.9 terrabyte computer vision dataset with real-world robot dynamic sequences paired with synthetic RGB-D, semantic, and instance segmentation maps for ML and VIO applications.
- Developed ARM Cortex-M4 firmware for ultra high-speed control of ESCs, realtime interpolated angular motor feedback using custom IR sensors, I/O safety watchdogs, muxing with I2C robotic commands, and realtime human e-stop and override controls transmitted via radio.
- Implemented a high-speed Velodyne 16 Puck LiDAR driver for use with LCM. Co-developed a framework of perception algorithms for rapid

control of an autonomous car under skidding conditions using LIDAR data.

- Developed and publically released a high-speed (360FPS) universal OptiTrack motion capture driver for use with ROS and LCM, written in C++ for performance with extra optimizations for network latency compensation.
- Experimented with using semantic segmentation maps to improve VIO robustness through improved rejection of outlying feature tracks in the high-speed VIO frontend before reaching the VIO non-linear optimizer.
- During real-world agile flight $(\geq 30.9mph, \geq 2.8G)$ position estimates are obtained from motion-capture camera images are rendered at $\leq 180\,\mathrm{Hz}$ and live streamed to the UAV – allowing for testing of VIO algorithms under arbitrary environment conditions.

MIT CSAIL[1] — Summer-Fall 2015

Researcher in Natural Language Processing (NLP)

- Developed deep learning models for natural language processing to automatically cluster and analyze large-scale medical research datasets.
- Created a classical NLP algorithm that extracts salient conclusions from unstructured oncology papers and synthesizes relevant results to user queries, reduced average reading load on a researcher from 8 pages per article to 2-5 sentences per article.

Akamai Technologies — Summer 2014

Server Platform QA Engineering Intern

- Engineered a server stress-testing tool 2.7+ times more powerful than Akamai's prior tool; reduced costs of Akamai's QA team by 8x for large-scale production tests.
- Architected a fast, dynamic file generation server that is flexible, easier to use, and 3x faster than Akamai's previous system.

Achievements & Honors

MIT Museum Exhibiter — 2015-2017

Creator of an on-going interactive robotics exhibit that interacts with $\sim$200K visitors/year.

- Co-created a popular, interactive robotic prosthetic arm exhibit at the MIT Museum using custom fabricated 0402 SMD PCBs, ARM Cortex-M4 firmware, and capacitive touch sensors.
- Robot on display year-round with more than 200k visitor interactions and 0% interactivity downtime.

MIT Hack Med Prize — Nov 2014

Product designer of *Opi-pal*, a prize-winning hardware solution for treating opioid overdose.

- Delivered a functional CAD prototype and a looks-like 3D printed prototype in less than 24 hours; completed prototype product placed top 3 in MIT Hacking Medicine's H^3 Hackathon.

Patents

"AI-DRIVEN AUGMENTED REALITY MENTORING AND COLLABORATION"
SRI International Inc, 2024.

"SYSTEMS AND METHODS FOR VERTICAL FARMING"
Oishii Farm Corporation, 2025.

"SYSTEM AND METHOD FOR DETECTING AND MANAGING PESTS"
Oishii Farm Corporation, 2025.

"POLLINATION SYSTEM"
Oishii Farm Corporation, 2025.

Selected Publications (656+ cit.)

IJRR '19: Int'l Journal of Robotics Research
"The Blackbird UAV Dataset".
Antonini, **Guerra***, Murali, Sayre-McCord, and Karaman.*

IROS '19: Int'l Conf. on Intelligent Robots & Sys.
"FlightGoggles: Photorealistic Sensor Simulation for Perception-driven Robotics using Photogrammetry and Virtual Reality".
***Guerra**, Tal, Murali, Ryou, and Karaman.*

ACC '19: American Controls Conference
"Perception-aware trajectory generation for aggressive quadrotor flight using differential flatness".
*Murali, Spasojevic, **Guerra**, and Karaman.*

[1] CSAIL: Computer Science Artificial Intelligence Laboratory

*Both authors contributed equally to this work.

MIT '19: Masters of Engineering Thesis

"Photorealistic Sensor Simulation for Perception-driven Robotics using Virtual Reality".

Guerra.

ISER '18: Int'l Symp. on Experimen'l Robotics

"The Blackbird Dataset: A large-scale dataset for UAV perception in aggressive flight".

*Antonini, **Guerra**, Murali, Sayre-McCord, and Karaman.*

ICRA '18: Int'l Conf. on Robotics & Automation

"Visual-inertial navigation algorithm development using photorealistic camera simulation in the loop".

*Sayre-McCord, **Guerra**, Antonini, Arneberg, Brown, Cavalheiro, Fang, Gorodetsky, McCoy, Quilter, Riether, Tal, Terzioglu, Carlone, and Karaman.*

ISEC '17: Integrated STEM Education Conf.

"Project-based, collaborative, algorithmic robotics for high school students: Programming self-driving race cars at MIT".

*Karaman, Anders, Boulet, Connor, Gregson, **Guerra**, Guldner, Mohamoud, Plancher, Shin, et al.*

IJID 2017: Int'l Journal of Infectious Diseases

"Planning an innovation marathon at an infectious disease conference with results from the International Meeting on Emerging Diseases and Surveillance 2016 Hackathon".

*Ramatowski, Lee, Mantzavino, Ribas, **Guerra**, Preston, Schernhammer, Madoff, and Lassmann.*