Rishikesh Bhyri

+1 (716)-923-5597 | rbhyri@buffalo.edu | rishi1134.github.io | linkedin.com/in/rishikesh-bhyri | github.com/rishi1134

Education

University at Buffalo, The State University of New York Buffalo, NY
Master of Science in Computer Science ***– CGPA: 3.95/4.0*** *Aug 2024 - Jun 2026*

- Thesis: Enhancing Spatial Reasoning in Vision Transformers for High-Fidelity Object Counting and Verification.
- Advisor: Prof. Junsong Yuan
- Coursework: Machine Learning, Parallel and Distributed Processing, Reinforcement Learning, Deep Learning

Vellore Institute of Technology Chennai, India
Bachelor of Technology in Electronics and Communication Engineering ***– CGPA: 9.48/10.0*** *Jul 2017 – Jun 2021*

- Capstone: Logo Detection and Analysis in Images.
- Advisor: Prof. Sathiya Narayanan. S

Publications and Patents

- **Rishikesh Bhyri** et al., "Chain-of-Look Spatial Reasoning for Dense Surgical Instrument Counting" IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2026. **(Accepted)** (Paper)
- **Rishikesh Bhyri**, Nan Xi, Junsong Yuan, "Structured Object Counting with Visual Chain Reasoning" European Conference on Computer Vision (ECCV), 2026. **(In Review)**
- **Rishikesh Bhyri**, Nan Xi, Junsong Yuan, "Medical Video Diffusion Reasoning" European Conference on Computer Vision (ECCV), 2026. **(In Review)**
- Multi-Agent Computer Vision system and methods for automated object counting, classification and tracking. USPTO Application No. 63/851,812. Status: Patent Pending

Research Experience

Graduate Research Assistant Buffalo, NY
University at Buffalo (Advisor: Prof. Junsong Yuan, Dr. Peter C W Kim) *Oct 2024 – Present*

- **High-Density Object Counting (CountGD Extension):** Engineered a novel counting framework extending CountGD (GroundingDINO, Swin-B, BERT). Integrated class-specific learnable tokens and designed a domain-specific loss function to enhance spatial reasoning in dense clusters, achieving a state-of-the-art **Mean Absolute Error (MAE) of 0.88**.
- **Surgical Instrument Instance Segmentation:** Developed a robust pipeline for low-to-medium density instrument segmentation. Benchmarked and fine-tuned Mask R-CNN, Mask2Former, and SAM 3, achieving **90% mAP** on the surgical dataset.
- **Density-Adaptive Mobile Deployment:** Built an Android application for real-time inference that utilizes a custom region proposal network to dynamically route image patches to either the segmentation model (low density) or the counting model (high density).
- **Performance Optimization:** Achieved a peak end-to-end latency of **0.32s** (vs. 25.12s human counting) on mobile hardware. Reduced manual effort by **99%** while maintaining high-fidelity detection visualizations.

Computer Vision Intern Remote
PlaEye LLC (Portland, OR) (Supervisor: Mr. Venkat Yellepeddy) *Jan 2021 – Jun 2021*

- **Logo Detection and Analysis in Images:** Developed a **mobile-based brand analytics system** for **logo detection and classification** using the LogoDet-3k dataset, optimized for arm64-v8a devices with ArmNN. Addressed challenges in dataset imbalance and diverse image features by implementing scaled loss functions and class augmentations. (Link)
- **Automatic Calibration of Tennis Court Images:** Created a novel algorithm for automated tennis **court calibration using Homography and Cross ratios** to overcome **single-view calibration limitations** arising from elusive feature points in occluded views. (Link)

Professional Experience

Machine Learning Intern San Jose, CA
Mercedes-Benz R&D North America *May 2025 – Aug 2025*

- Optimized inference by exploring **CUDA Graphs** with **Nsight Systems**, trimming CPU launch overhead and **cutting latency 20%**.
- Drove **4× faster inference** via INT8 **implicit and explicit quantization.**
- Built a C++/CUDA utility to inspect TensorRT engine weights and automate network-pruning sweeps.
- Implemented **unified-memory zero-copy** paths and custom GPU pools, **lowering peak GPU usage by 18%** and eliminating two memcpy calls per frame.
- Proved bit-exact **determinism** at runtime & layer level; stress-tested execution contexts to meet **ASIL-D safety**.
- Designated as **Primary Inventor** for **2 patents** (currently in filing stage).

Software Engineer Chennai, India
Citi *Jul 2021 – Jul 2024*

- Co-owned and led the development of a web-based end-to-end tool initiative for data creation and conditioning, reducing the time and effort required for data handling. Engineered the tool using ReactJS and NodeJS.
- Developed API layers in .NET using minimal-API, exposing Selenium-C# automation functionalities and enabling remote headless execution which **decreased data creation time by 5x**, reduced cross-team dependency and **turnaround time by 80%**.

Recent Projects

Masked-ResShift: Pixel-Level Residual Shift for Image Inpainting | *Paper* Jan 2025 – May 2025

- Optimized diffusion-based image inpainting by implementing novel strategies, including Masked-Variance Diffusion (MVD) and Exact Masked-Markov Diffusion (EMMD), achieving faster inference times while maintaining high image quality.

Traffic light automation using Reinforcement Learning | *Link* Jan 2025 – May 2025

- Designed a dynamic traffic control system using SUMO-RL to minimize wait times for emergency vehicles.
- Trained and benchmarked agents using **SARSA, DQN, Double DQN, and A2C** algorithms, demonstrating superior performance over static baselines in reducing emergency transit delays.

MindVault: AI-Powered Bookmark Organizer | *UBHacking'24 Best AI/ML Project* Nov 2024

- Chrome extension using BeautifulSoup, NLTK, Sentence Transformers, and DBSCAN for context-aware bookmark clustering.

Open-Source Contribution | *Segment Anything Model 3 (SAM 3)* Nov 2025

- Assisted users by answering queries regarding fine-tuning bugs and model configuration.
- Provided debugging support that helped close 3 GitHub issues: #200, #260, and #286.

Academic Service

Peer Reviewer
Machine Vision and Applications (MVA) Journal, Springer *Feb 2026 – Present*

Graduate Student Assistant – Grader Buffalo, NY
University at Buffalo (Supervisor: Dr. Asif Imran) *Jan 2025 – May 2025*

- Evaluated programming assignments, quizzes, and coursework for **CSE574: Introduction to Machine Learning**.

Teaching Assistant Chennai, India
Vellore Institute of Technology (Supervisor: Dr. Vetrivelan P) *Jan 2020 – May 2020*

- Evaluated programming assignments and quizzes for **ECE2005: Probability Theory and Random Process**.

Technical Skills

Languages: Python, C++, OpenMP, MPI, CUDA
Tools: PyTorch, TensorRT, ArmNN, ONNX, OpenCV, Scikit-Learn, Git, CMake, GCP, Slurm, Docker, LaTex
Concepts: Deep Learning, Computer Vision, Vision Transformers, Diffusion Models, Object Detection, Data Structures, Linear Algebra, Object Oriented Programming, VLM, PEFT

Honors & Awards

- **Best AI/ML Project, UBHacking'24** (Nov 2024): Won the 'Best AI/ML Project' award for developing an AI-powered, context-aware bookmark manager extension for Chrome.
- **Top Performance Rating, Citi**: Awarded a top rating (1 out of 4) for innovation in developing automated reporting tools and dashboards.
- **Academic Meritorious Award**: Achieved 4th overall department rank in B.Tech Electronics and Communication Engineering (2017–2021 batch).