RISHIKESH BHYRI

+1 (716)-923-5597 | bhyririshikesh@gmail.com | GitHub | LinkedIn | Website | Buffalo, NY

EDUCATION

Masters in Computer Science and Engineering, University at Buffalo, United States Aug 2024 - Jun 2026
Research with specialization in Computer Vision and Deep Learning **GPA - 3.95/4.0**
Courses: Machine Learning, Parallel and Distributed Processing, Reinforcement Learning, Deep Learning

Bachelor of Technology in Electronics and Communication Engineering Jul 2017 - Jun 2021
Vellore Institute of Technology, India **CGPA - 9.48/10**

TECHNICAL SKILLS

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

WORK EXPERIENCE

Research Assistant, Visual Computing Lab, University at Buffalo Oct 2024 - Present

- Built **high-density object-counting** model using **GroundingDINO**, **Swin-B**, and **BERT** with **cross-attention fusion**; improved accuracy via **semantic prompt tuning** and custom loss **(MAE < 1)**.
- Trained **ResNet & DINOv3**-based **MaskRCNN/Mask2Former** for low-density segmentation **(MAE < 0.5)**.

Machine Learning Intern, Mercedes-Benz R&D, San Jose 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 use 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**.

Software Engineer - II, Citi, India 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%**.

Computer Vision Intern, PlaEye LLC (Portland, OR), Remote Jan 2021 - Jun 2021

- 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)
- 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)

PATENTS AND PUBLICATIONS

- WACV 2026 Conference (Accepted): Chain-of Look Spatial Reasoning for Dense Surgical Instrument Counting
- Multi-Agent Computer Vision system & methods for automated object counting, classification and tracking. USPTO Application No. 63/851,812. Status: Patent Pending.

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.

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

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