Ritesh Gangnani

Head of AI | Founding ML Engineer | Generative AI Expert

ritesh.gangnani@gmail.com

+91-7503214123

New Delhi, India

17/09/1996

https://www.linkedin.com/in/riteshgangnani

https://github.com/riteshgangnani10

6+ years of AI/ML expertise from computer vision research to **deploying 250+ production GenAI APIs**. **Co-creator of VoltaML** (4-6x Stable Diffusion speedup). **Led end-to-end AI product development** at Segmind, architecting proprietary tools and scaling inference infrastructure.

Skills

Gen AI & Models
SDXL, SSD-1B, ControlNet, LoRA, Inpainting, Flux, Stable Diffusion

Optimization
VoltaML, TensorRT, TorchScript, xformers, CUDA

Infrastructure & MLOps
Docker, FastAPI, Kubernetes, AWS, GCP, RunPod, REST APIs

Computer Vision
GANs, ResNet, U-Net, YOLOv5, CNNs, Image/Video Segmentation

Tools & Libraries
PyTorch, ComfyUI, TensorFlow, OpenCV, HuggingFace

Education

M.Sc. Computer Science
South Asian University
2017 – 2019

B.Sc. Computer Science
University of Delhi
2014 – 2017

Professional Experience

Head of AI / Lead Engineer
Segmind Inc.
02/2023 – Present | Delhi, India

Led complete AI product development lifecycle for Segmind's GenAI API platform:

- **Architected and deployed 250+ production GenAI APIs** - responsible for all model hosting, optimization, and deployment infrastructure
- **Conceptualized and built multiple proprietary IP tools** including face swapping, virtual try-ons, product photography, and interior design solutions
- **Integrated VoltaML backend achieving 4-6x inference speedup** - directly contributed to platform scalability and cost efficiency
- **Built Pixelflow:** Modular ComfyUI-based pipeline system enabling rapid deployment of new AI workflows
- **Scaled production infrastructure** using Docker, FastAPI, Triton, RunPod, AWS, GCP
- **Drove product strategy** by identifying market opportunities and translating them into technical solutions

VoltaML (Open Source Project) — Co-Creator & Contributor
04/2022 – 03/2023

- **Co-initiated and architected VoltaML** - open-source project focused on accelerating Stable Diffusion inference
- **Engineered core optimization engine** using TensorRT, TorchScript, ONNX, and CUDA achieving **4-6x performance improvements**
- **Achieved significant community adoption** - 1.2K+ GitHub stars, became foundational to Segmind's backend architecture
- **Established industry benchmark** for Stable Diffusion inference optimization

Machine Learning Scientist → ML Manager
Onward Assist
10/2019 – 01/2023 | Hyderabad, India

Built end-to-end computer vision pipelines for histopathology AI applications:

- **Developed comprehensive CV architecture** using GANs, U-Net, ResNet, and CNNs for medical image analysis including:
  - Stain normalization algorithms
  - Advanced image segmentation models
  - Tumor classification systems
- **Deployed production-ready medical AI tools** using Docker & Kubernetes, integrated into clinical trial workflows
- **Promoted to ML Manager (2020)** - led technical team, managed client-facing medical AI tools, collaborated directly with pathologists
- **Delivered enterprise solutions** for Telepath Dx and GE Healthcare, presented at HIMSS 2022

Publications

Improving Classification of Lymph Node Histopathology Patches Using Semi-Supervised Classification-GAN (SSC-GAN) - (Poster)
Nvidia GTC - 2020

Semi-supervised Multi-category Classification with Generative Adversarial Networks
PReMI - 2019