Upscaling for Flood Resilience: A Benchmarking Study

Jooho Kim[1], Mitchell Horn[2], and Yuming Han[3]

[1] Institute for Disaster Resilient Texas (IDRT), Texas A&M University, College station, TX
[2] Yale School of the Environment, Yale University, New Haven, CT
[3] Department of Electrical and Computer Engineering, Texas A&M University, College station, TX

Jooho.kim@tamu.edu

Introduction

High-resolution imagery is essential for disaster response and recovery, where clear visuals guide evacuation routes, flood mapping, and infrastructure repair. Without reliable data, responders risk delays and misallocation of resources. In Texas alone, more than 500 under-resourced communities lack updated floodplain information and monitoring systems, leaving them dependent on low-quality community-shared photos or videos. Enhancing these images directly supports faster, more accurate decision-making during floods and other hazards.

Challenge: Real-ESRGAN is a state-of-the-art AI model for restoring and enhancing degraded images. Running inference on a single dataset can require more than a full day of compute time, with sustained GPU utilization. This high cost makes it difficult to integrate into time-sensitive disaster workflows, where rapid results are critical. While powerful accelerators—ranging from high-memory GPUs to next-generation processors and wafer-scale engines—are available, most prior studies evaluate model design or hardware specifications in isolation. This leaves a gap in understanding how to best deploy these models at scale.

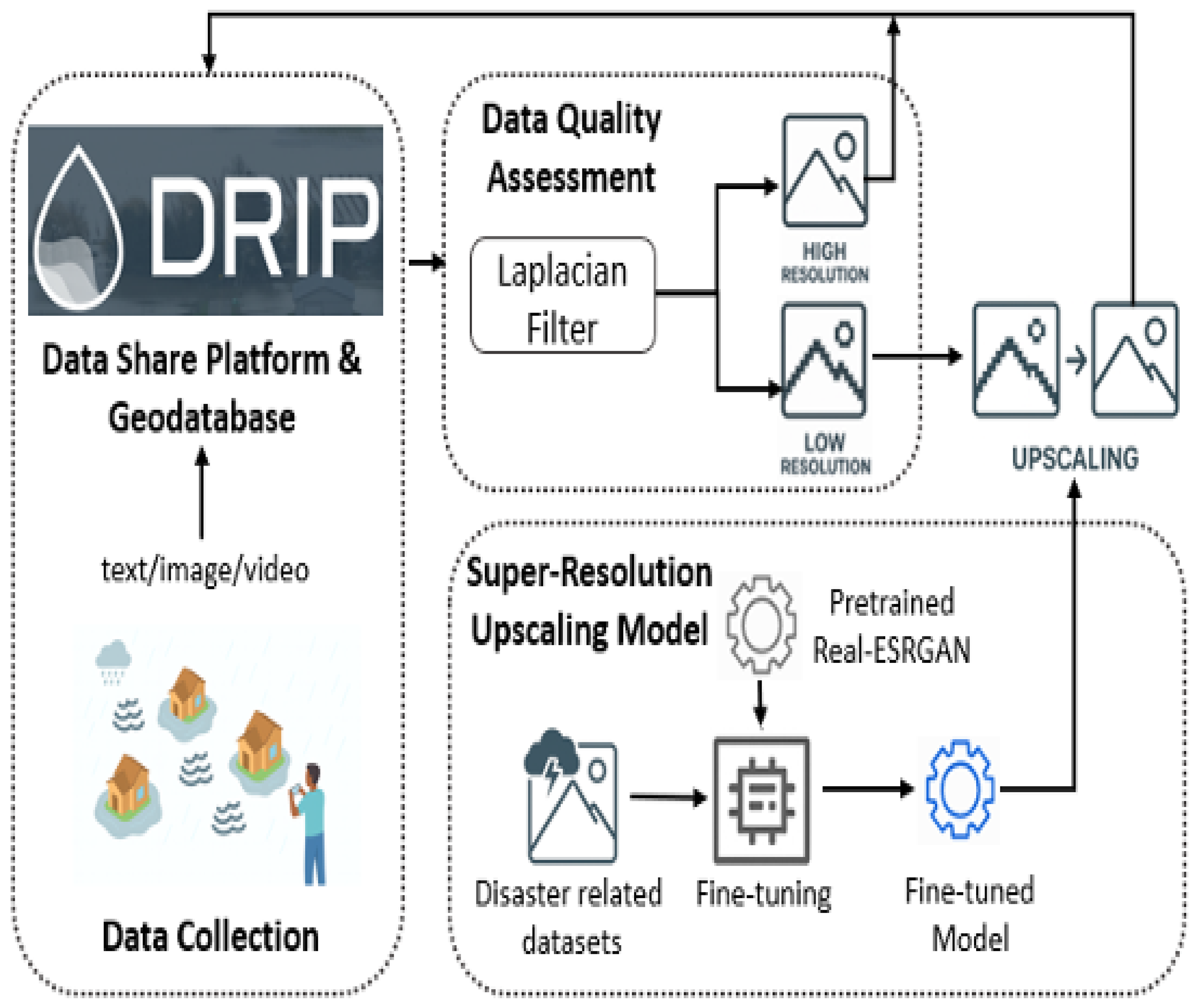
Approach/Impact: Develop a portable benchmarking pipeline for Real-ESRGAN that runs seamlessly across different accelerator platforms. Measure performance tradeoffs in runtime, throughput, memory use, and efficiency under varying dataset sizes. Use both image and video datasets relevant to disaster scenarios to ensure results reflect real-world conditions. The outcome is faster delivery of clearer, more actionable imagery to disaster-impacted communities—supporting situational awareness, resource allocation, and long-term resilience planning.

Methods

Dataset: Data in the form of text, images, and videos is first collected and stored in a shared platform and geodatabase (DRIP). Then, a data analysis step using a Laplacian filter separates high- and low-resolution inputs; the high-resolution data are returned to the database, while the low-resolution inputs are processed through the super-resolution upscaling pipeline.

Sample Images

Model Structure

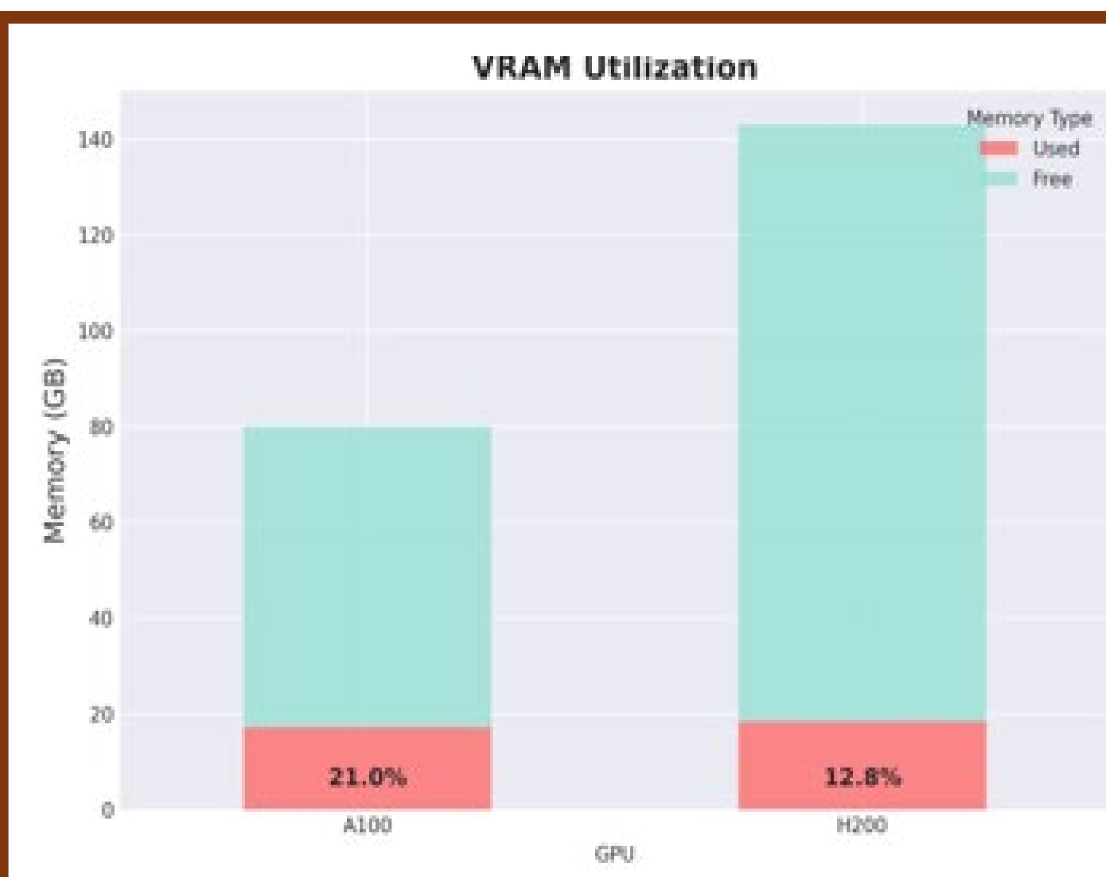


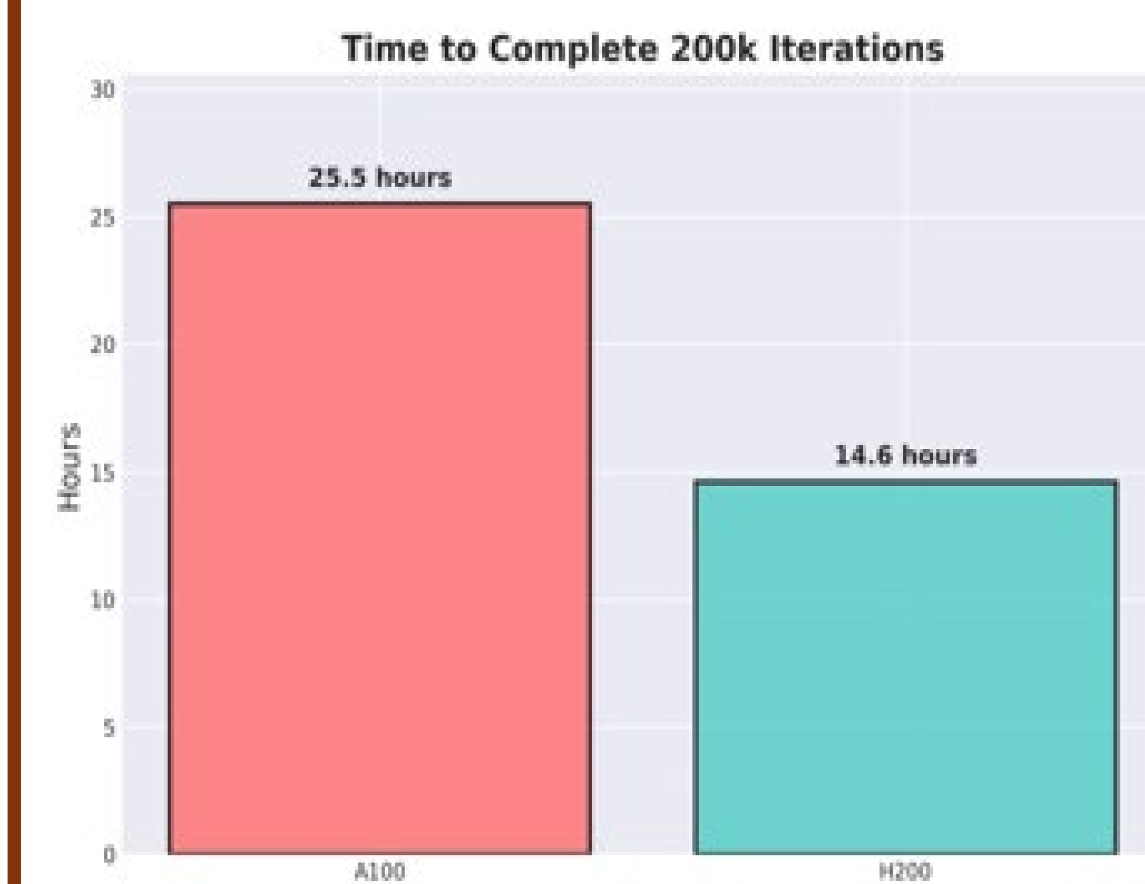
GPU Architectures

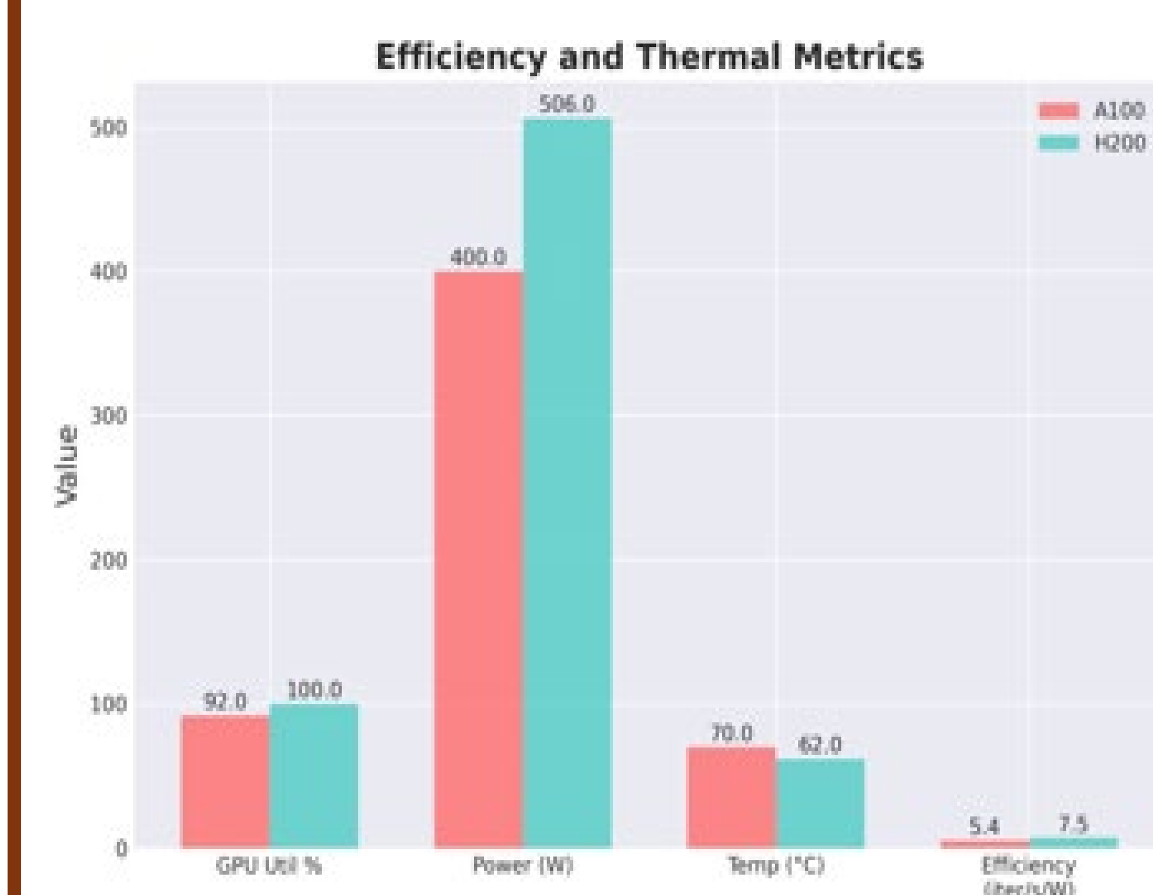
	PROFESSIONAL	ENTERPRISE	ENTERPRISE	RESEARCH
Model	L40S	A100	H200	Bow-2000
Architecture	Ada Lovelace	Ampere	Hopper	Bow IPU (3nm)
MEMORY				
Capacity	48 GB HBM	80 GB HBM2	141 GB HBM3	14.4 GB*
Bandwidth	864 GB/s	2.04 TB/s	4.8 TB/s	47.5 TB/s*
PERFORMANCE				
FP16 (TFLOPS)	733*	624	1,979	250*
SYSTEM SPECS				
Power (TDP)	350W	400W	700W	1500W
Interconnect	PCIe 4.0	NVLink 3	NVLink 4	IPU-Link 2Tbps

* Bow-2000 at full saturation 4 IPU processors, 16 tiles

Results

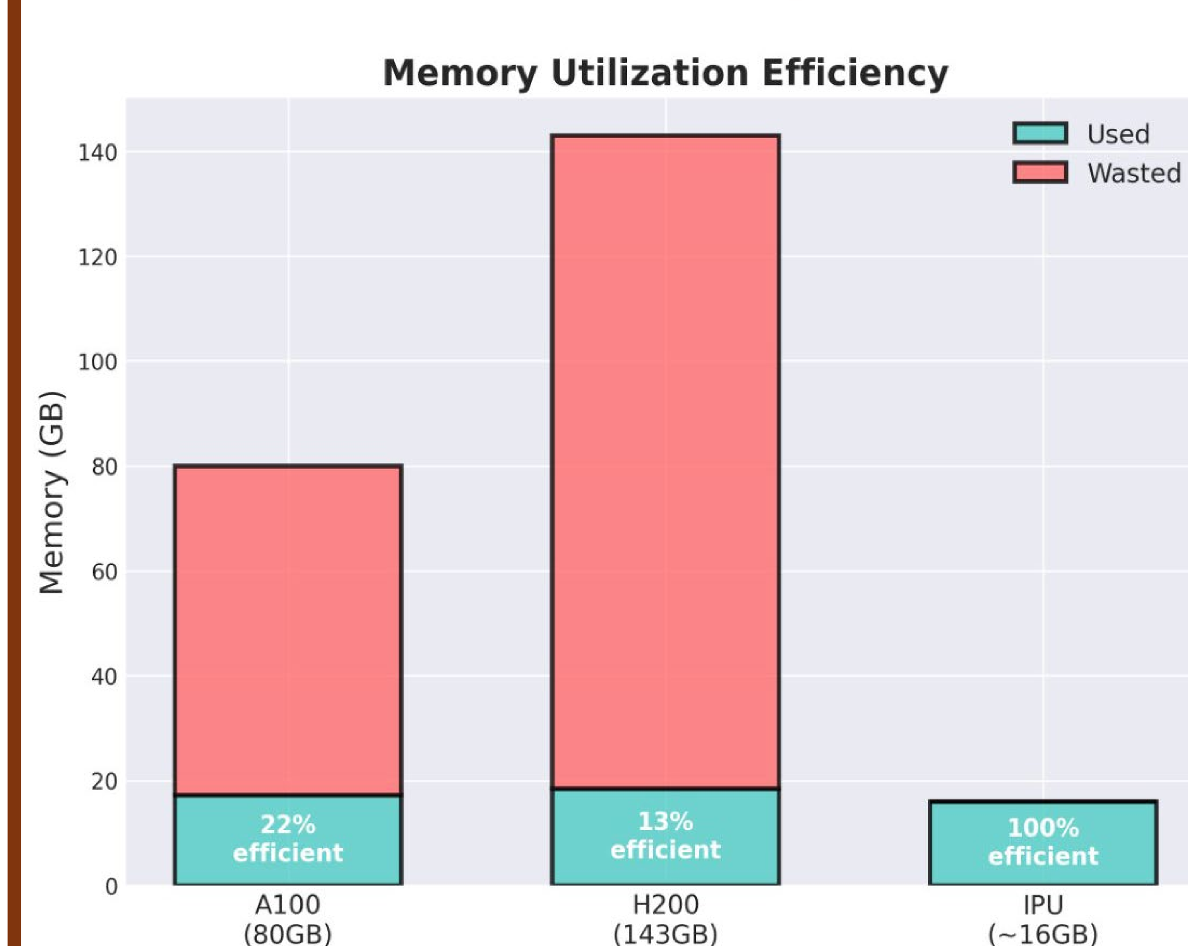






H200 not only trains faster and runs cooler but also delivers better efficiency per watt, despite drawing more total power. The expanded VRAM and improved throughput make it a stronger choice for large-scale AI model training.

Future/in-development



IPU's offer an interesting architecture for this type of high-throughput and low-memory training. While performance gains were significant in jumping to higher memory bandwidths in the H200 compared to A100, there is significantly more wasted VRAM. At full saturation of the Bow-2000 engine, there is even more throughput potential without waste.

References/Acknowledgements

- *ByteBoost '25 Cybertraining Program*
- *NSF Awards: 2320990, 2320991, and 2320992,*
- *ACES (Texas A&M), Bouchet (Yale), Delta (UIUC)*