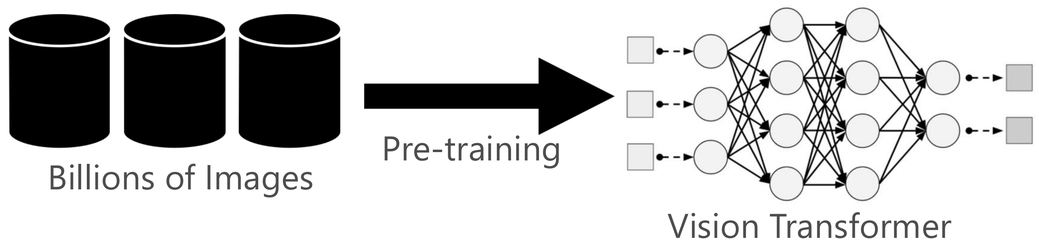
Towards High Performance Image-Free ViT Pre-training Using Tensor Based Fractal Generation

Edgar Josafat Martinez-Noriega[1], Truong Thao Nguyen[1], Jason Haga[1], Yusuke Tanimura[1], and Rio Yokota[2]
[1]The National Institute of Advanced Industrial Science and Technology (AIST), [2]Institute of Science Tokyo, Japan

Introduction

Pre-training Vision Transformers (ViTs) conventionally relies on large-scale image corpora, often comprising millions of samples.

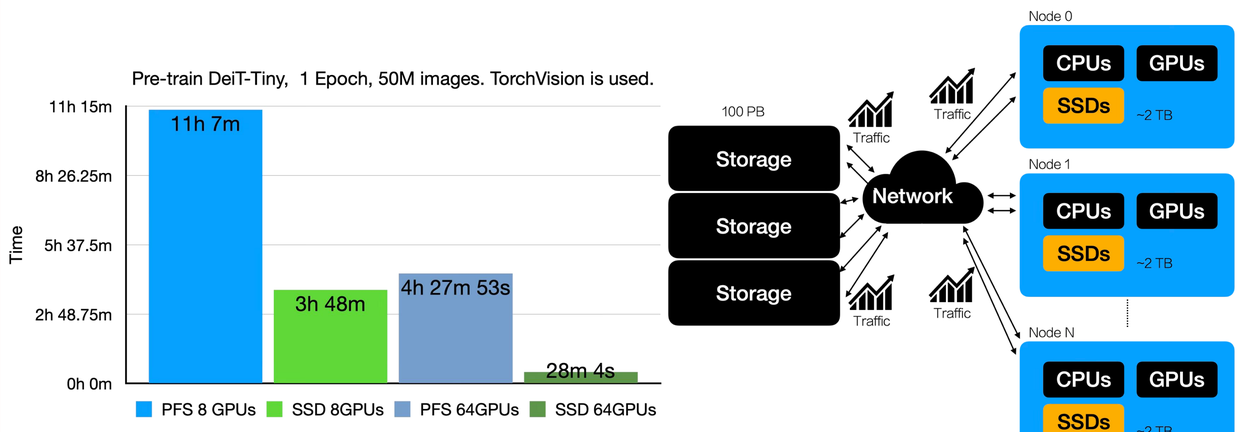


Real-world datasets are typically used for this purpose but present significant challenges, including high storage requirements, limited accessibility, and ethical risks.

Dataset	No. of Images	Open to public
CIFAR10	50 Thousand	Yes
ImageNet 1k	1.2 Million	Yes
ImageNet 21k	18.4 Million	Yes
JFT-300M	300 Million	No
JFT-3B	3 Billion	No

Challenges

However, as dataset sizes scale beyond couple of million images, I/O bottlenecks and CPU operations introduce significant delays in pre-training [2].



As dataset sizes scale to several million samples, I/O and CPU-bound routines become significant bottlenecks. Caching the dataset on local SSDs substantially reduces pre-training time, and this benefit persists even as the number of GPUs increases.

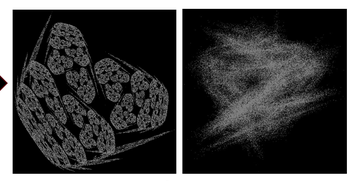
Method

Synthetic datasets have emerged as a promising alternative, particularly within the framework of Formula-Driven Supervised Learning (FDSL) [1], which employs mathematically defined fractals and complex geometric constructs for pre-training, followed by fine-tuning on downstream real-world tasks.

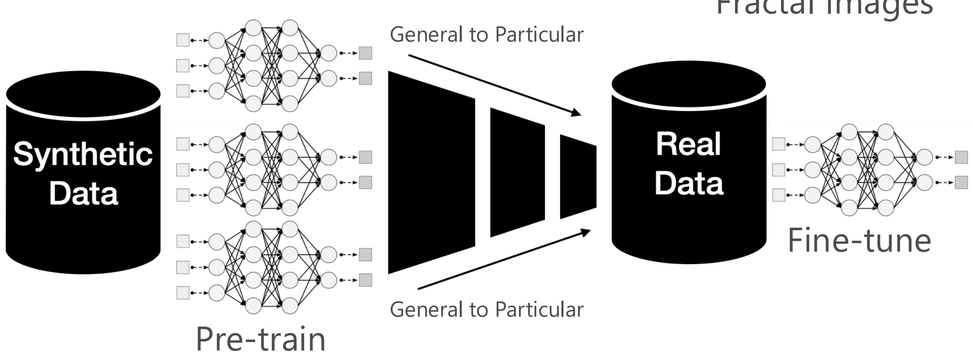
$$w_i(x;\theta_i) = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} x + \begin{bmatrix} e_i \\ f_i \end{bmatrix}$$

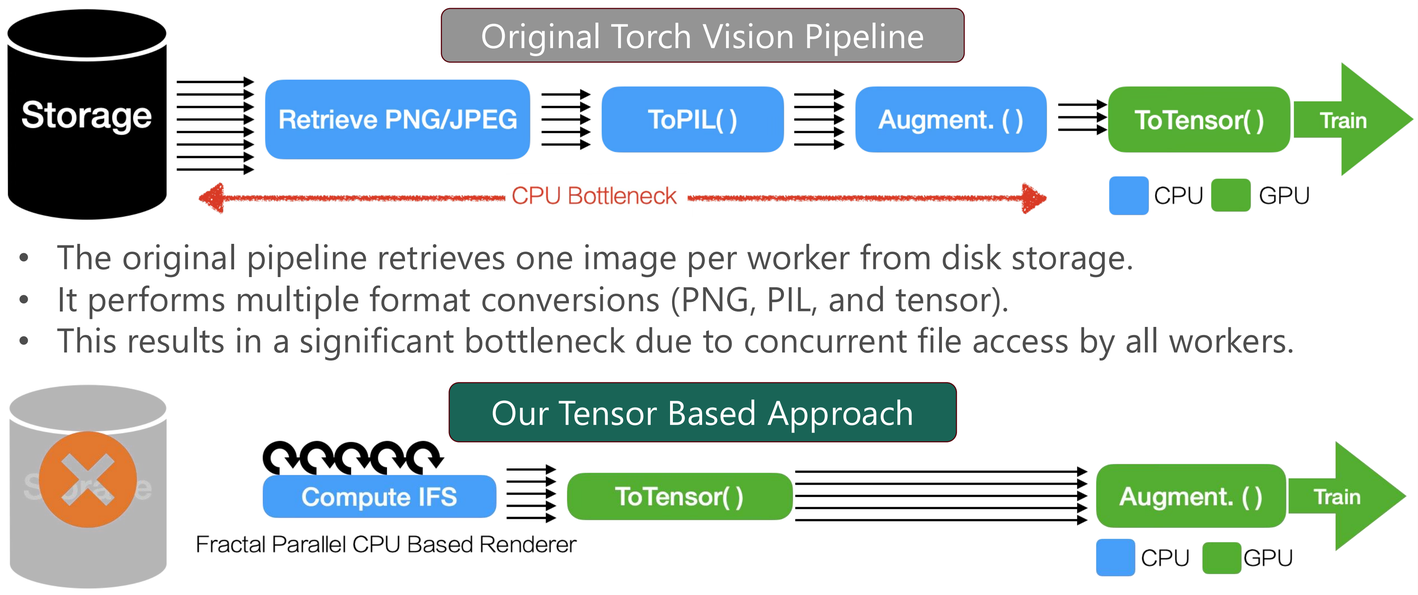
Iterated Function System (IFS)

Rendering



Fractal Images





- The original pipeline retrieves one image per worker from disk storage.
- It performs multiple format conversions (PNG, PIL, and tensor).
- This results in a significant bottleneck due to concurrent file access by all workers.

- We proposed a tensor-based fractal generation pipeline to minimize CPU involvement.
- We implement an IFS renderer in C/Python through a PyBind11 interface. The C backend receives the IFS parameters and generates the fractal images directly in memory.
- Our pipeline generates fractal images on the fly during data loading, reducing format conversions and I/O overhead while enabling high-throughput execution.

Preliminary Results

Experimental Set-up

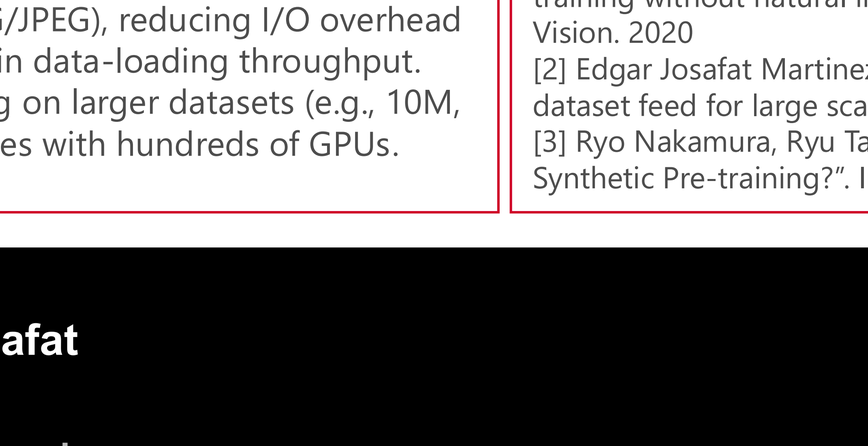
We follow Nakamura et al. [3], restricting the dataset to one fractal instance per class. All experiments were executed on the ABCI 3.0 with the following specs:

- CPU: Intel Xeon 8558, 48x2 Cores,
- Mem: 16x64 GB DDR5
- SSD: NVMe 3.2TB
- GPU: 8xH200 SXM 141GB

On the experiments: ABCI AI Bridging Cloud Infrastructure

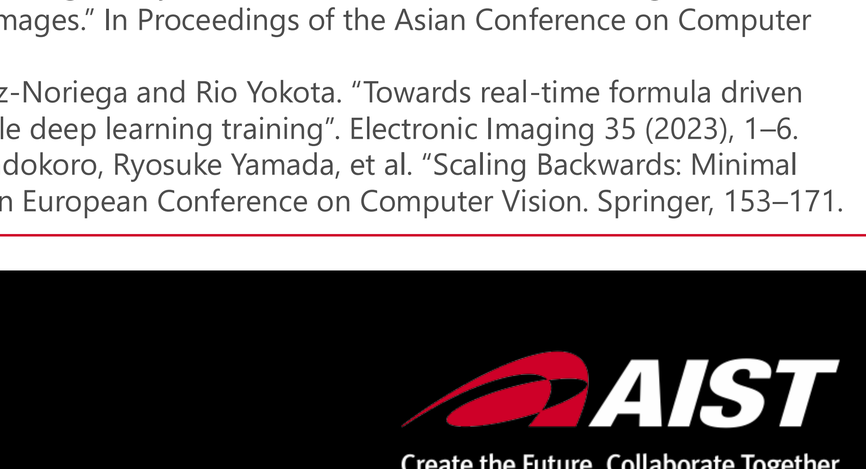
- 20 workers per GPU.
- Dataset: RT-FractalDB (1M images)
- Model: DeiT-Tiny.

Loading Time Profile



- We scale to a full eight-GPU node in distributed mode and vary the batch size (BS) to show performance.
- Our approach maintains nearly constant loading times across all BS, remaining below 20s due to efficient parallel rendering and direct tensor retrieval.
- TorchVision requires 1m 12s compared to only 13.7s using our approach.

Full Pre-training Time



- The TorchVision pipeline exhibits severe latency that is 4x larger than our approach when loading 1M images on a single. GPU.
- Our method completes a full epoch using DeiT-Tiny in 1m 4s, whereas the TorchVision requires 2m 4s.
- Our method can reduce the end-to-end training time by nearly 50%.

Conclusion & Future Work

- We propose a tensor-based fractal generation pipeline that eliminates disk access and image decoding (e.g., PNG/JPEG), reducing I/O overhead and achieving up to a 3.95x improvement in data-loading throughput.
- We plan to conduct more detailed profiling on larger datasets (e.g., 10M, 21M, and beyond), scaling to multiple nodes with hundreds of GPUs.

References

[1] Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto, Eisuke Yamagata, et al. “Pre-training without natural images.” In Proceedings of the Asian Conference on Computer Vision. 2020
[2] Edgar Josafat Martinez-Noriega and Rio Yokota. “Towards real-time formula driven dataset feed for large scale deep learning training”. Electronic Imaging 35 (2023), 1–6.
[3] Ryo Nakamura, Ryu Tadokoro, Ryosuke Yamada, et al. “Scaling Backwards: Minimal Synthetic Pre-training?”. In European Conference on Computer Vision. Springer, 153–171.

MARTINEZ NORIEGA Edgar Josafat
AIST, Tokyo Waterfront, Japan
Contact: edgar.martineznoriega@aist.go.jp

AIST
Create the Future, Collaborate Together