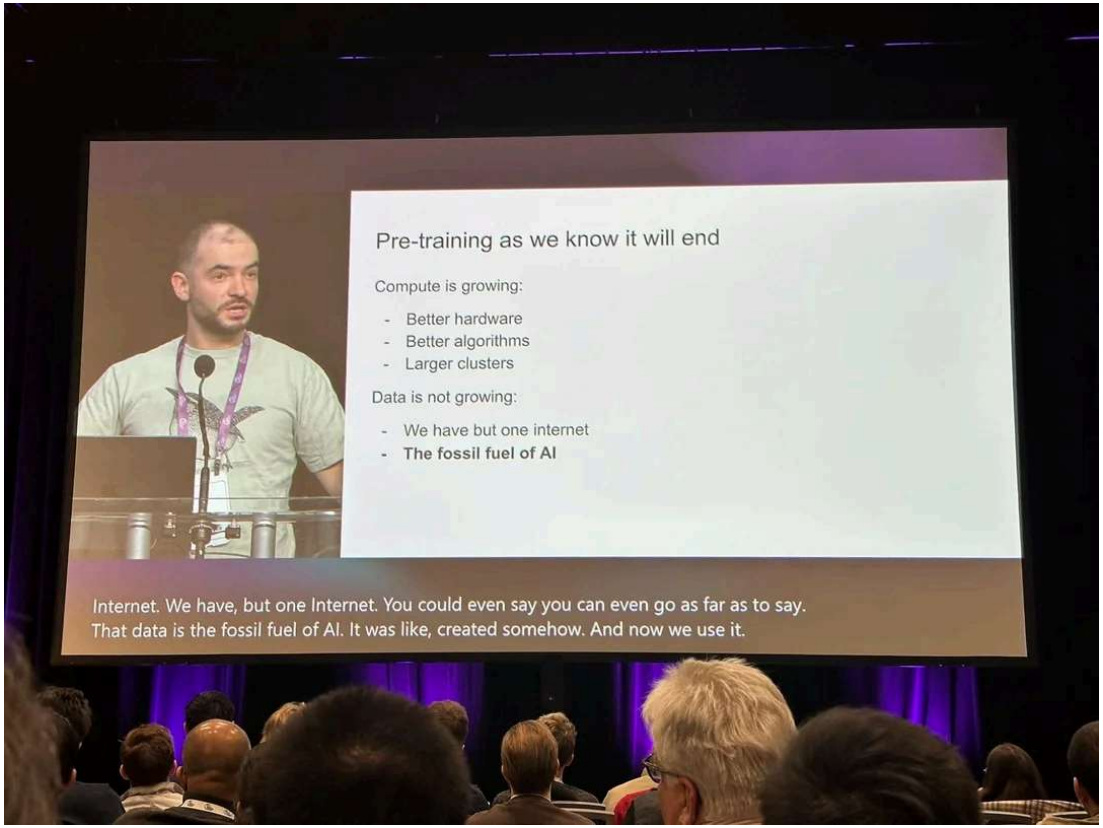
High-Efficiency Online Data Generation to Improve Pretraining Scaling Laws of Deep Networks

SIKE OGIEVA ‘25

COLLABORATING WITH

- **SUSHMIT CHAKMA (HAVERFORD COLLEGE)**
- **RANDALL BALESTRIERO (BROWN UNIVERSITY)**

The Data Scarcity Problem in AI



Inspiration: AutoAugment

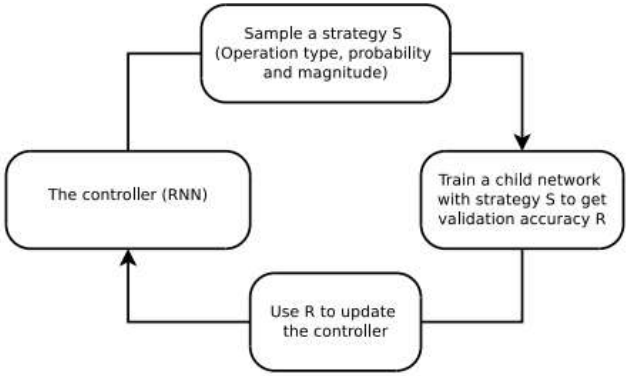


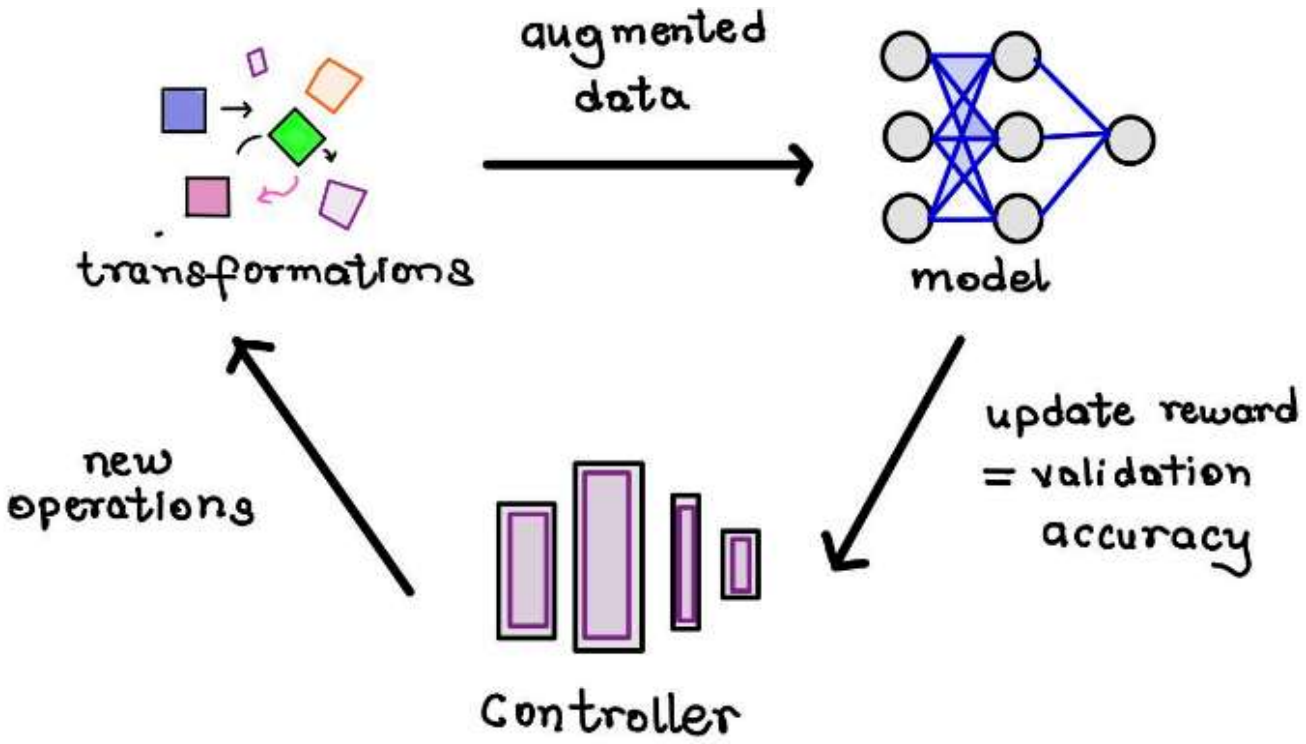
Figure 1. Overview of our framework of using a search method (e.g., Reinforcement Learning) to search for better data augmentation policies. A controller RNN predicts an augmentation policy from the search space. A child network with a fixed architecture is trained to convergence achieving accuracy R. The reward R will be used with the policy gradient method to update the controller so that it can generate better policies over time.

Dataset	GPU hours	Best published results	Our results
CIFAR-10	5000	2.1	1.5
CIFAR-100	0	12.2	10.7
SVHN	1000	1.3	1.0
Stanford Cars	0	5.9	5.2
ImageNet	15000	3.9	3.5

Table 1. Error rates (%) from this paper compared to the best results so far on five datasets (Top-5 for ImageNet, Top-1 for the others). Previous best result on Stanford Cars fine-tuned weights originally trained on a larger dataset [66], whereas we use a randomly initialized network. Previous best results on other datasets only include models that were not trained on additional data, for a single evaluation (without ensembling). See Tables 2,3, and 4 for more detailed comparison. GPU hours are estimated for an NVIDIA Tesla P100.

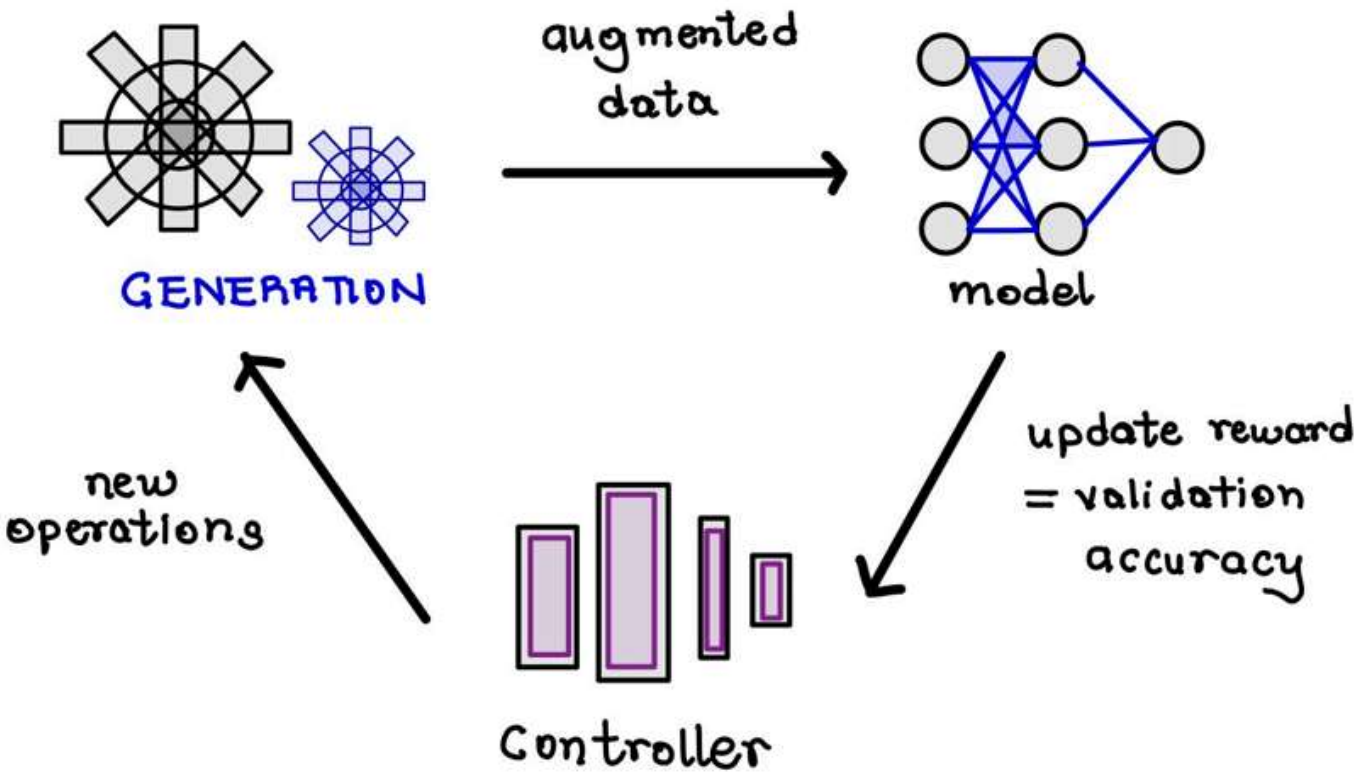
Cubuk et al, AutoAugment: Learning Augmentation Policies from Data, 2019

AutoAugment



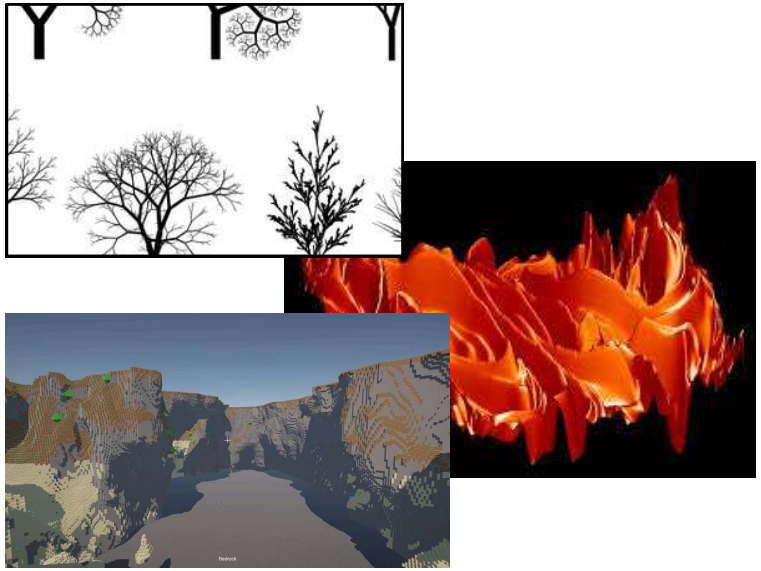
Cubuk et al, AutoAugment: Learning Augmentation Policies from Data, 2019

Our Framework



Procedural Image Generation

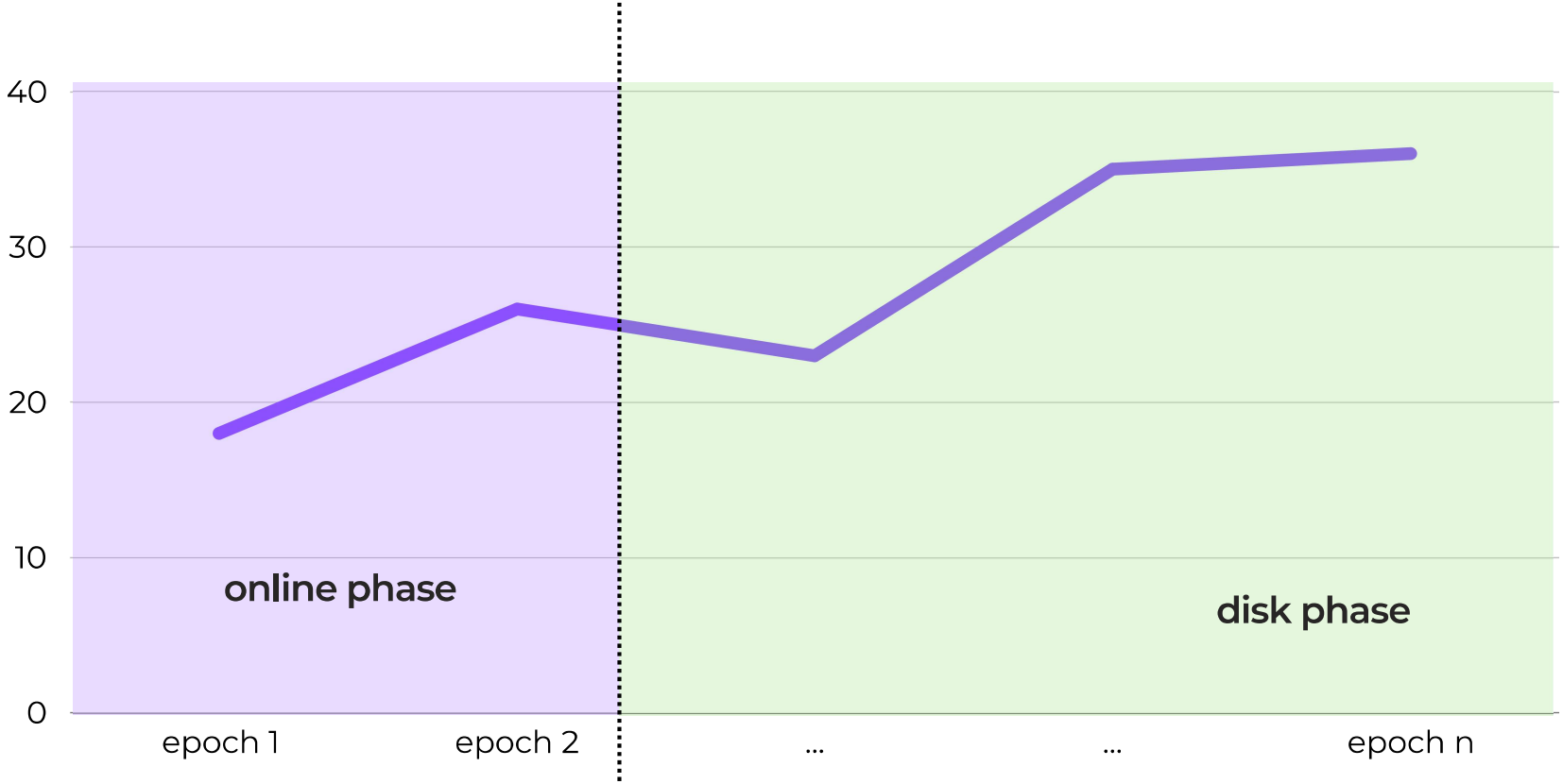
Infinigen, 2023



Small Open-Source Libraries

Last Recourse: Small GANs

Warm-Up Stage



Next Steps

1. Watch for developments
    a. On-the-fly Dataset Augmentation with Synthetic Data (Li et al, 2025)
2. Pure generation
3. Larger Datasets