Vision Backbone Efficient Selection for Image Classification in Low-Data Regimes

J. Guerin[1], **S. Bansal**[2], **A. Shaban**[3], **P. Mann**[4] & **H. Gazula**[5]
[1]Espace-Dev, IRD, Univ. Montpellier ; [2]Georgia Institute of Technology ; [3]Field AI ;
[4]Institute of Computing, Federal University of Rio de Janeiro ; [5]Massachusetts Institute of Technology
Contact: joris.guerin@ird.fr

1- Introduction

- Transfer learning crucial in low-data regimes.
- Backbone selection is often overlooked [1].

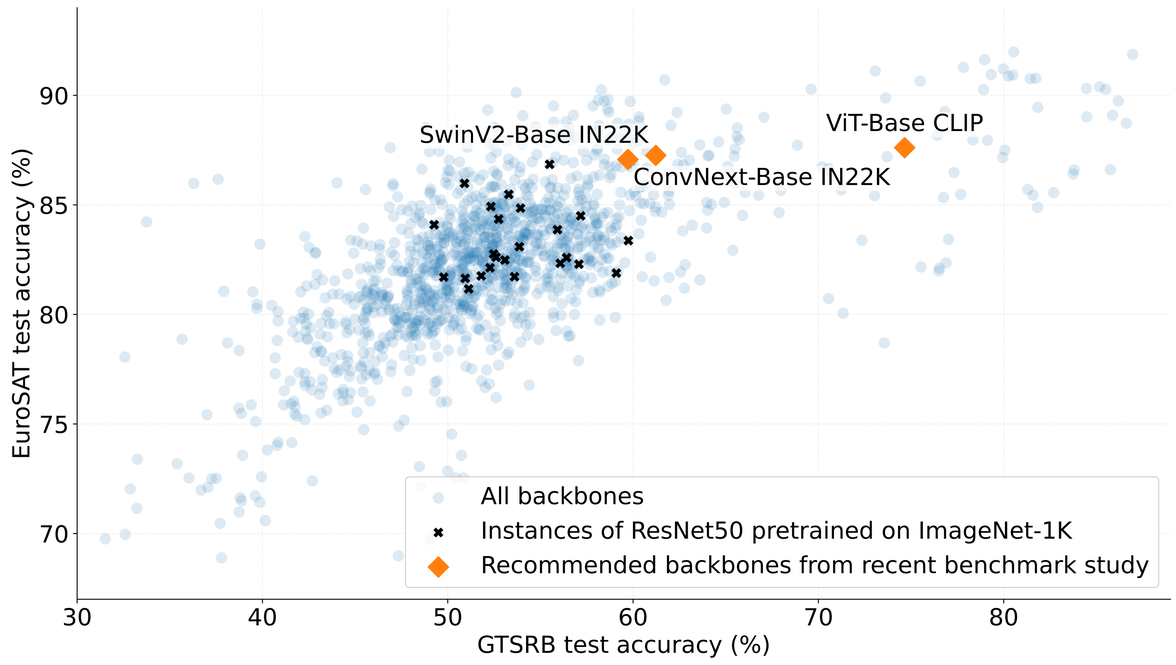


Fig. 1: Comparing backbones on 2 datasets. Logistic regression on frozen backbones, 10 samples / class. Recommended from [2].

Benchmark studies faces critical limitations:

- **Limited coverage:**
  - ∼50 models vs. thousands in DL libraries.
  - Benchmarks results quickly outdated.
- **Overemphasis on average performance:** Generic recommendations display suboptimal behavior (Fig. 1).
- **Neglecting Implementation Variability:** Benchmarks fail to capture variability among similar models (ResNet50 in Fig. 1).

Dataset specific backbone selection can yield significant performance gains with minimal additional effort compared to hyperparameter tuning or architectural modifications.

2- Problem Formulation

Notations

- Target task $\mathcal{T}$
  - Datasets ($\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}, \mathcal{D}_{\text{test}}$).
  - Metric ϵ: ϵ_{val} and ϵ_{test} are validation/test performance).
- Backbone pool $\mathcal{B} = \{b_1, \dots, b_N\}$. For each b, we train a light classifier (e.g., log. reg.) on features from b. We write $\epsilon_{\text{val}}(b)$ and $\epsilon_{\text{test}}(b)$.

Vision Backbone Selection

$$b^*_{\text{val}} = \arg\max_{b \in \mathcal{B}} \ \epsilon_{\text{val}}(b),$$

proxy for $b^*_{\text{test}} = \arg\max_b \epsilon_{\text{test}}(b)$.

Exhaustive search time:

$$t = \sum_{b \in \mathcal{B}} \tau(b)$$

$\tau(b)$ time to assess b (download, extraction, training, validation) – impractical for large $\mathcal{B}$.

VIBES: Vision Backbone Efficient Selection

Trade a bit of optimality for speed via two levers:

- Fast approximate evaluation: compute $\tilde{\epsilon}_{\text{val}}(b)$ with cost $\tilde{\tau}(b) \ll \tau(b)$.
- Optimized sampling: evaluate only a subset, using an ordering π over $\{1, \dots, N\}$ to prioritize promising models.

A VIBES solution is defined by $\tilde{\epsilon}$ and π. For time budget $t_{\max}$, the selected backbone $\hat{b}$ is:

$$\hat{b} = \arg\max_{i \in \{1,\dots,k\}} \tilde{\epsilon}_{\text{val}}(b_{\pi(i)}),$$

where k is the largest integer in $\{1, \dots, N\}$ such that $\sum_{i=1}^{k} \tilde{\tau}(b_{\pi(i)}) \leq t_{\max}$.

Evaluation: Backbone Selection Efficiency Curves (BSEC)

3- Approaches

Fast approximate evaluation: Replace logistic regression with faster learning-free models.

- Nearest Centroid
- KNN (K=5)

Optimized sampling:

- Random (baseline)
- ↑ Model complexity
- ↓ Model complexity
- ↓ Pretraining dataset size
- Dataset cycling

4- Experiments

Datasets: CIFAR10, GTSRB, Flowers102, EuroSAT

Low-data protocol: N=10 train examples/class; full test splits. Additional N in Appendix.

Backbone pool: 1,322 backbones from `timm`, spanning ∼0.5M to >1.2B params.

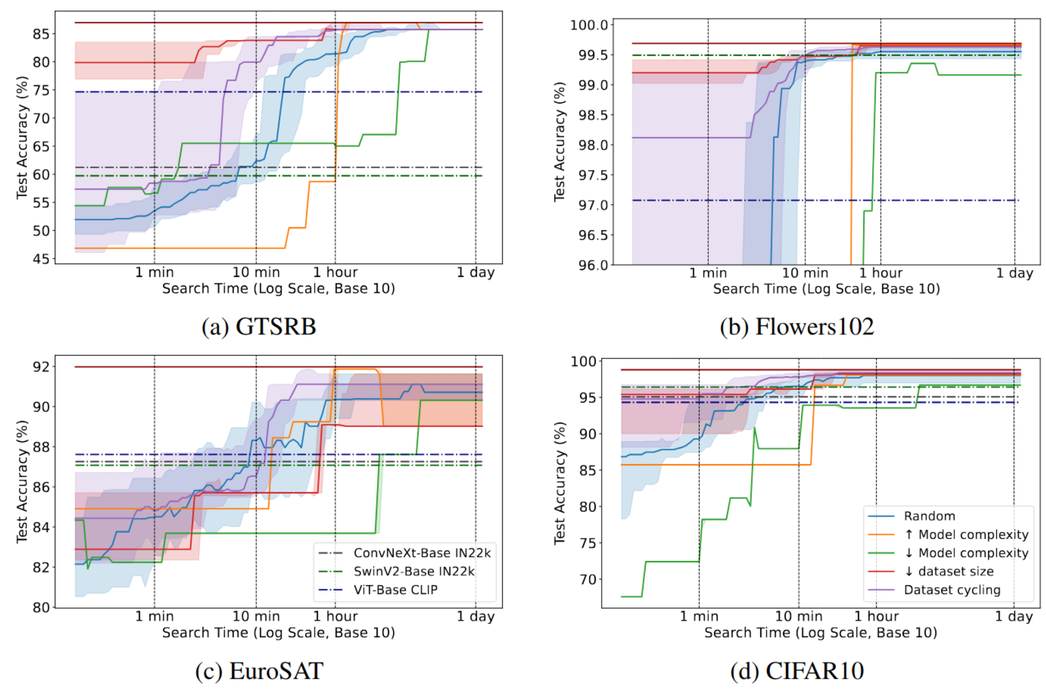
Hardware: NVIDIA RTX A5000 (24 GB).

5- Results

Fast approximate evaluation: No significant improvement

- Training << download/extraction

Optimized sampling strategies:



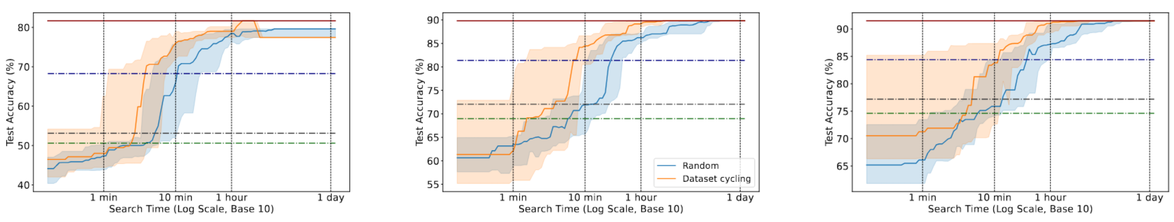
(a) GTSRB (b) Flowers102

(c) EuroSAT (d) CIFAR10

- **Dataset-specific vs. general-purpose.**
  - **10 min:** basic VIBES beats most general-purpose backbones.
  - **1 h:** VIBES beats all generic backbones.
  - Dataset-specific selection >> one-size-fits-all approaches.
- **Sampling strategy comparison.**
  - **By model complexity:** consistently *worse* than random.
  - **By pretraining data:** Outperform random. Domain relevance is key.

6- Varying Data Scracity

Additional experiments on GTSRB. Dataset cycling (best approach) vs. random sampling (baseline) under varying levels of data scarcity.



(a) $N = 5$ sample per class (c) $N = 20$ sample per class (d) $N = 40$ sample per class

Consistent performance advantage of dataset cycling demonstrating the robustness of our sampling strategy recommendation.

7- References

[1] J. Guérin et. al., "Combining pretrained cnn feature extractors to enhance clustering of complex natural images," *Neurocomputing*, vol. 423, pp. 551–571, 2021.
[2] M. Goldblum et. al., "Bob: A large-scale comparison of pretrained models across CV tasks," *Neural Information Processing Systems*, vol. 36, 2024.