

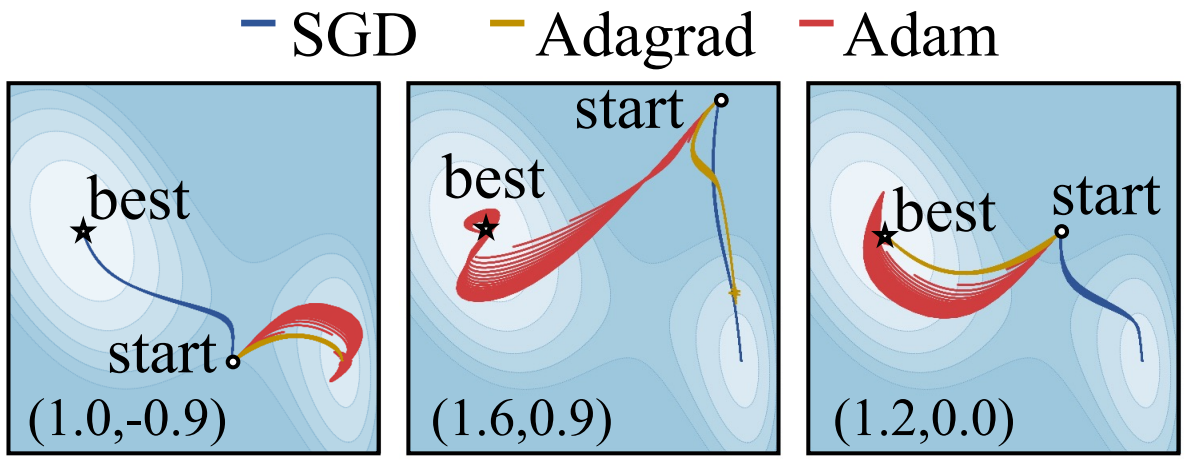
Towards Dynamic Interleaving Optimizer

Yile Chen, Zeyi Wen, Jian Chen, Jin Huang

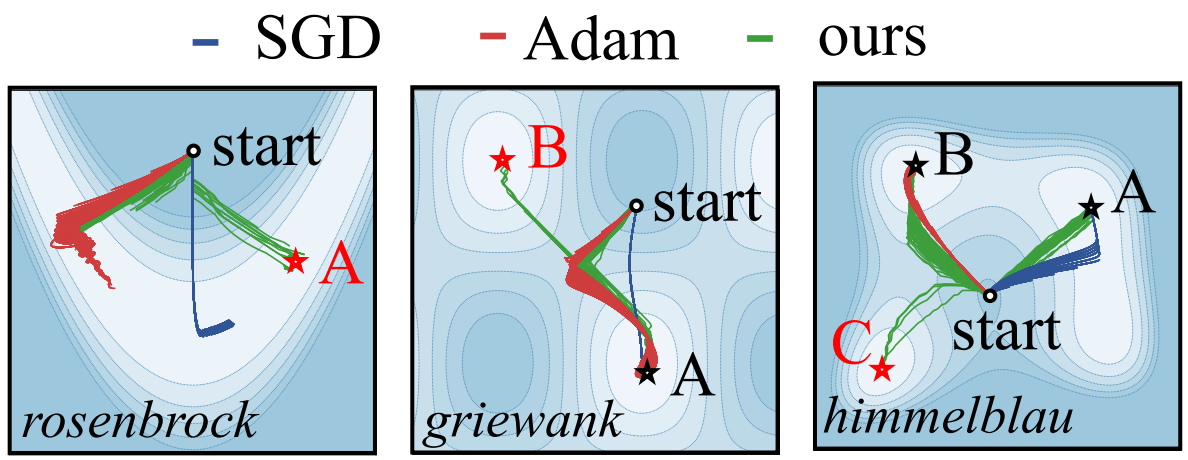
South China University of Technology
The Hong Kong University of Science and Technology (Guangzhou)
The Hong Kong University of Science and Technology
South China Normal University

Motivation

- Different optimizers are suitable for different training states.
- Transitioning between optimizers during training can enhance both convergence speed and model generalization.
- We propose Dynamic Optimizer Interleaving Training (DOIT), a novel framework that adaptively selects the suitable optimizer during training.



Optimal optimizers vary across different initial states.



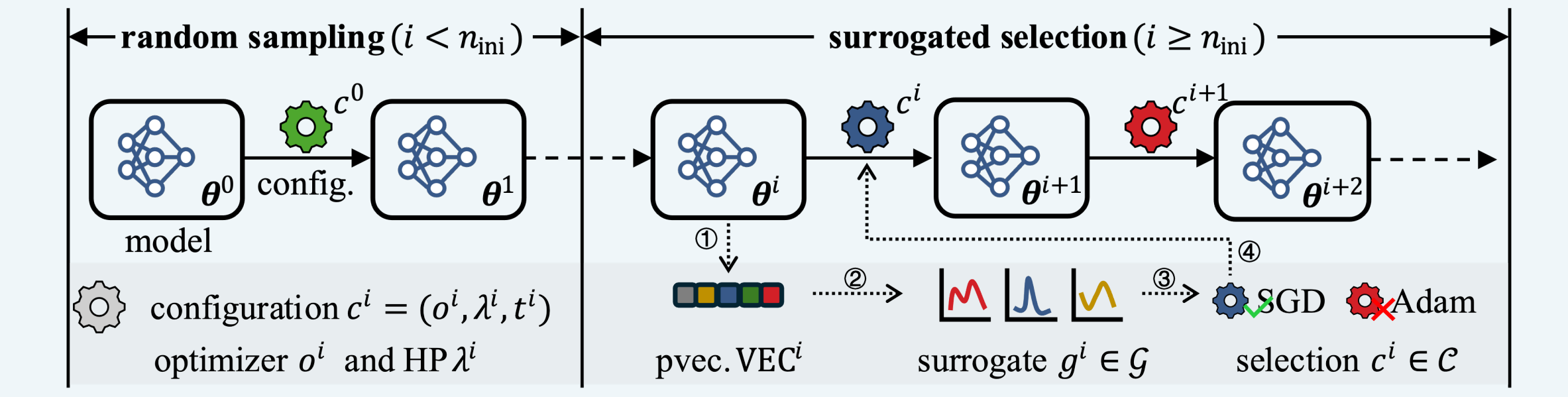
DOIT performs well across diverse parameter landscapes.

Our Method: surrogate model

- **Weighted random sampling:** collect diverse experience trajectories.
- **Surrogated selection:** predict performance of candidate optimizers.

  Input. Key model parameters VEC^i + HP λ^i.

  Output. Performance score $s = \tanh\big((\mu_\Delta + \Delta_{\mathrm{UPPER}} + \Delta_{\mathrm{LOWER}})/3 + \alpha\sigma_\Delta\big)$.



Training process of our proposed DOIT.

Our Method: acquisition function

- Consideration of variance.

$$ACQ(s_\mu, s_\sigma) = s_\mu + \alpha s_\sigma.$$

- Consideration of transferability.

$$ACQ(s_\mu, s_\sigma, \omega_t) = s_\mu + \alpha(1 - \omega_t)s_\sigma.$$

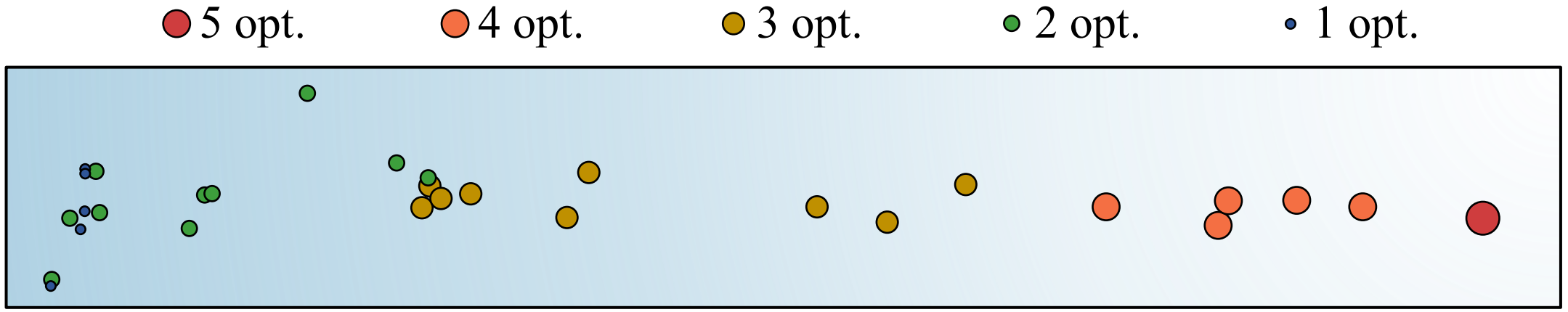
where ω_t is the transferability score.

- Consideration of training process.

$$e = \text{sigmoid}(s_\mu + \alpha(1 - 2^{-\lfloor i/n \rfloor} \cdot \omega_t)s_\sigma).$$

Our Method: surrogate model

- **Accuracy.** Achieving superior accuracy by unlocking novel optimization paths through multi-strategy synergy.



- **Time Cost.** predict performance of candidate optimizers.

	forward	gradient	optimizer	transferability	surrogate	acquisition
FLOPs	2.2×10^{16}	4.4×10^{16}	$0.2 \sim 2.0\times10^{13}$	2.2×10^{13}	5.3×10^{14}	1.0×10^{14}
Proportion	33.3%	66.7%	quite small	0.3‰	8.0‰	1.5‰

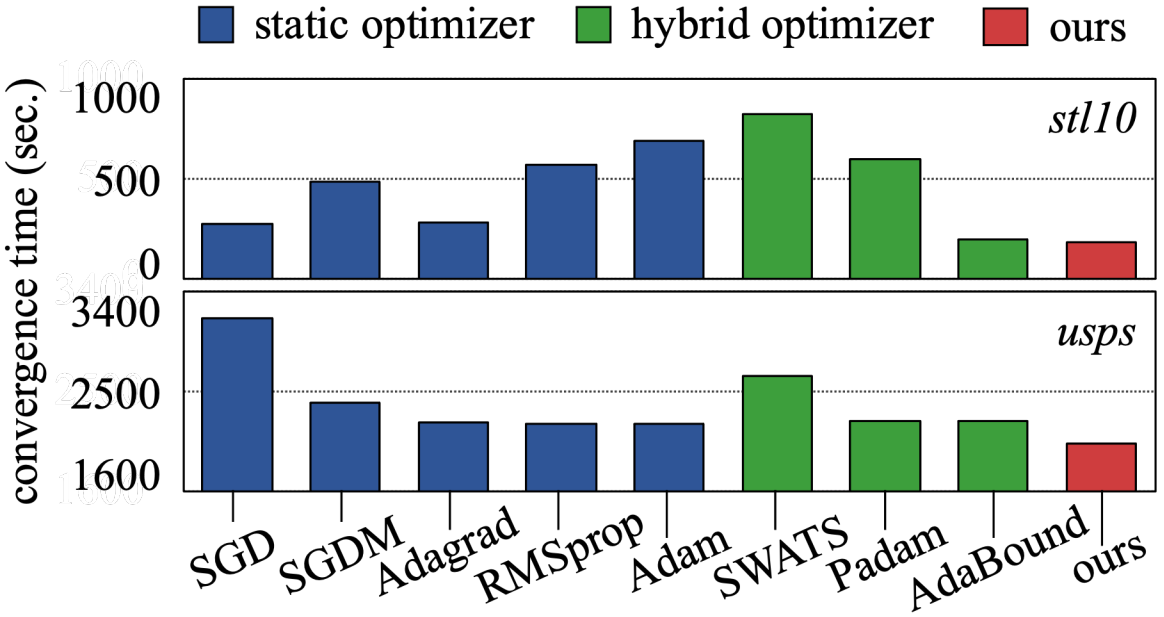
Experiment Results

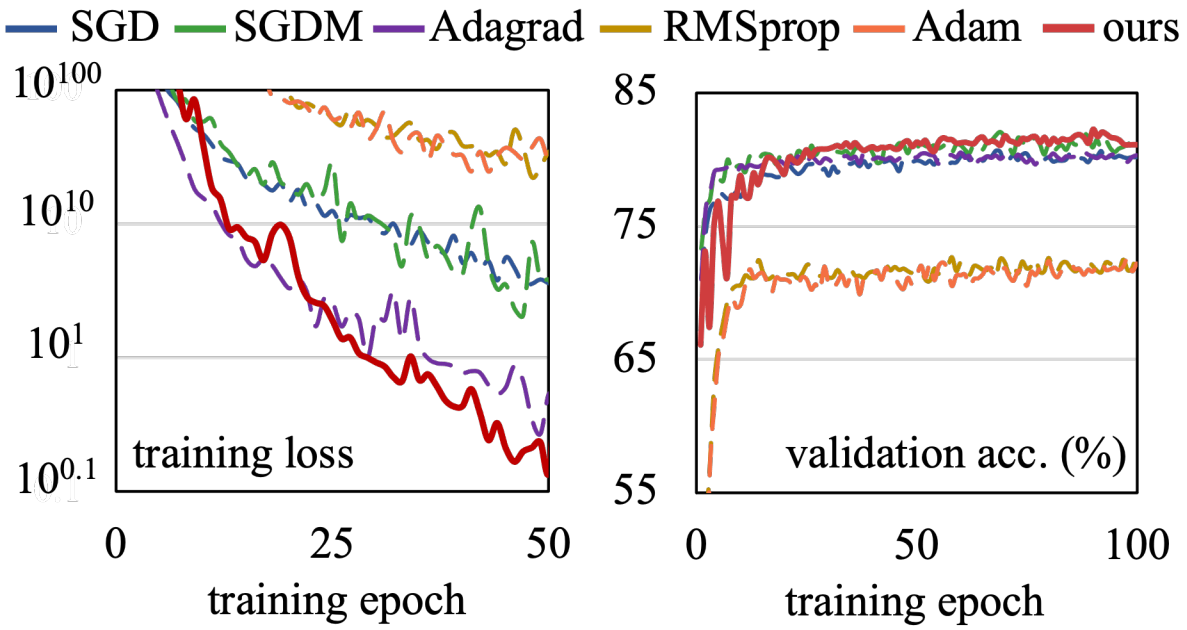
- Enhanced Accuracy & Stability.
  - 1%- 3% improvement across most benchmarks.
  - Lower variance, ensuring more robust training sessions.

	full training					PEFT			
	usps	mnist	stl10	cifar10	imagenet	usps	stl10	mrpc	qqp
SGD	97.46±0.70	99.50±0.04	97.83±0.22	97.53±0.03	78.73±0.38	94.42±0.21	97.75±0.16	85.21±0.35	82.13±0.52
SGDM	97.68±0.11	99.65±0.04	96.61±0.04	97.58±0.04	78.36±0.12	95.67±0.14	98.37±0.10	85.54±0.69	83.30±0.63
Adagrad	93.40±0.04	98.24±0.50	78.66±2.54	60.95±0.62	77.59±0.03	95.37±0.21	98.34±0.09	84.94±0.59	83.47±0.79
RMSprop	95.25±0.04	98.14±0.09	88.62±4.45	78.09±0.92	74.01±0.08	94.64±0.53	97.91±0.07	84.09±0.76	82.09±0.63
Adam	93.26±0.78	99.01±0.08	82.26±1.16	75.23±0.92	73.96±0.11	94.47±0.49	98.36±0.03	86.52±0.71	82.27±0.71
SWATS	94.00±1.23	98.73±0.13	88.03±0.44	66.15±4.74	76.93±0.19	95.12±0.21	98.38±0.10	86.27±0.62	80.79±0.81
Padam	97.58±0.11	99.66±0.04	90.81±0.06	96.03±0.03	77.47±0.16	95.72±0.42	98.38±0.10	80.64±0.32	73.04±0.91
AdaBound	87.64±0.84	97.54±0.13	86.33±2.39	70.91±4.35	75.93±0.13	95.42±0.11	98.30±0.14	68.38±0.59	78.64±0.49
ours	**97.81**±0.21	**99.71**±0.02	**98.21**±0.19	**98.04**±0.03	**79.98**±0.01	**96.12**±0.10	**99.01**±0.09	**87.99**±0.13	**85.57**±0.14

Experiment Results

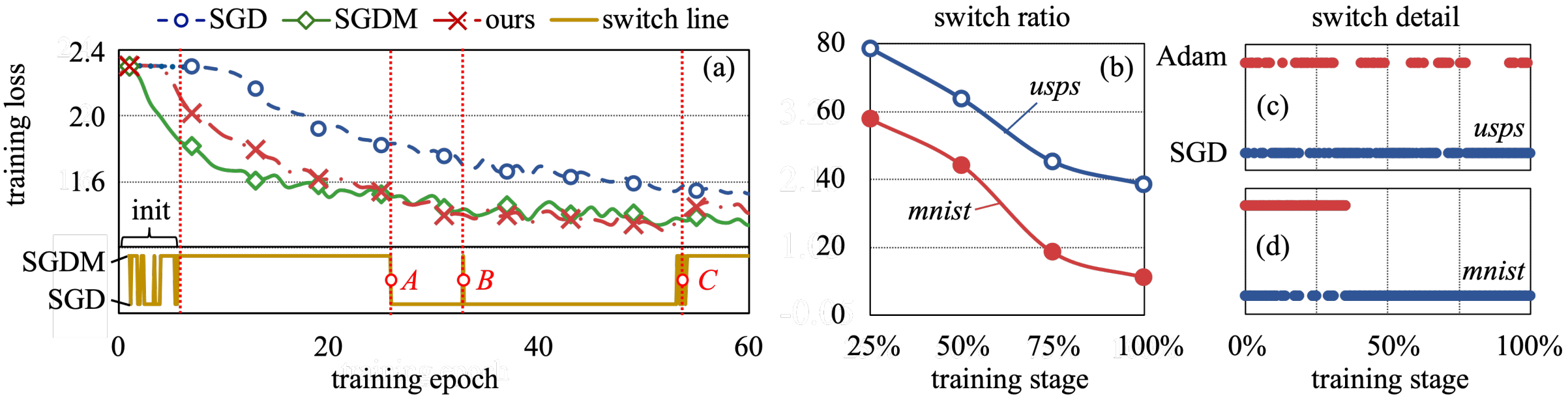
- Efficient Convergence.
  - Speedup in convergence with minimal computational overhead.





Selection Preference

- *As training progresses, the selection becomes more stable.*
- *In early phase, there is a tendency to select faster optimizers (e.g., Adam), later tends to select more stable optimizers (e.g., SGD).*
- *DOIT's selection preference is more apparent in a large dataset.*





THANKS !

Contact: Yile Chen
Email: jireh.x6@gmail.com