Faculty of Computer Science

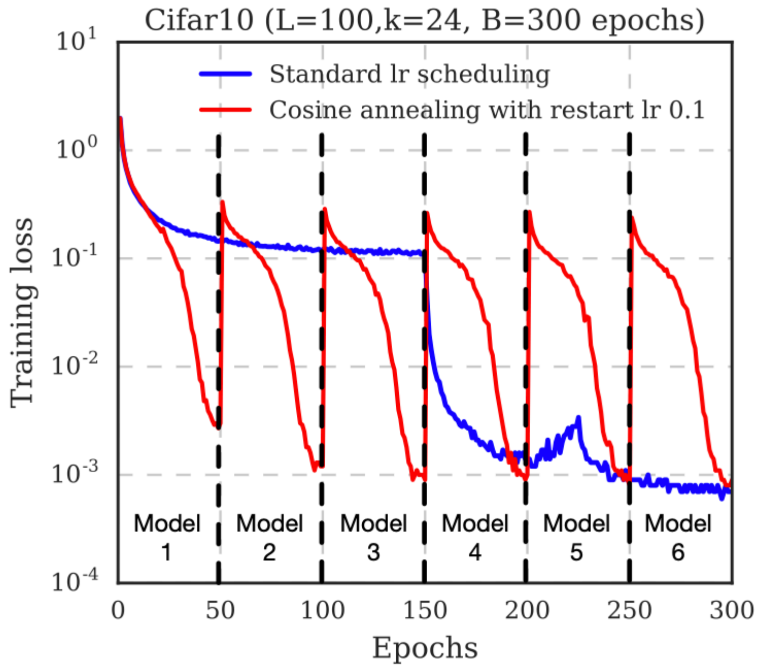
Applied Mathematics &
Computer Science

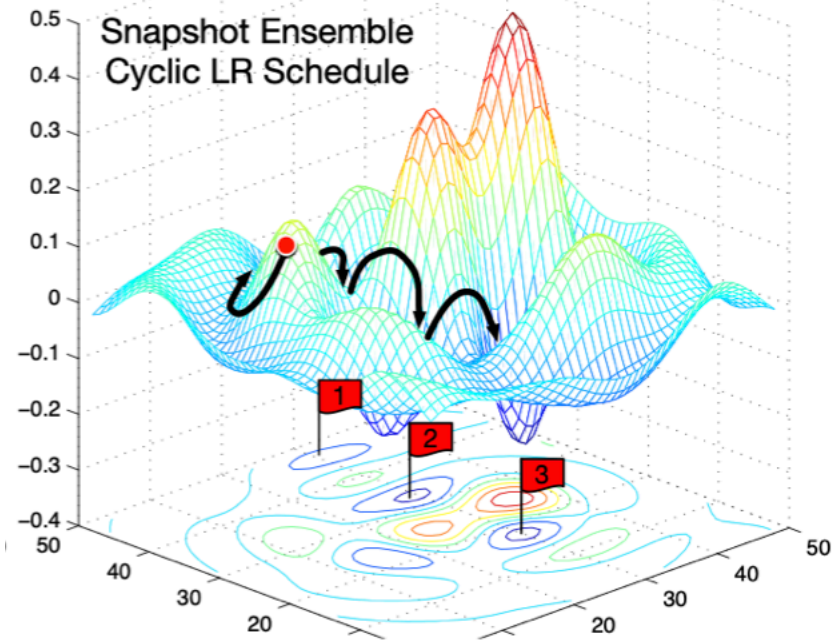
Moscow 2024

Analysis of Neural Networks Internal Representations During Transfer Learning

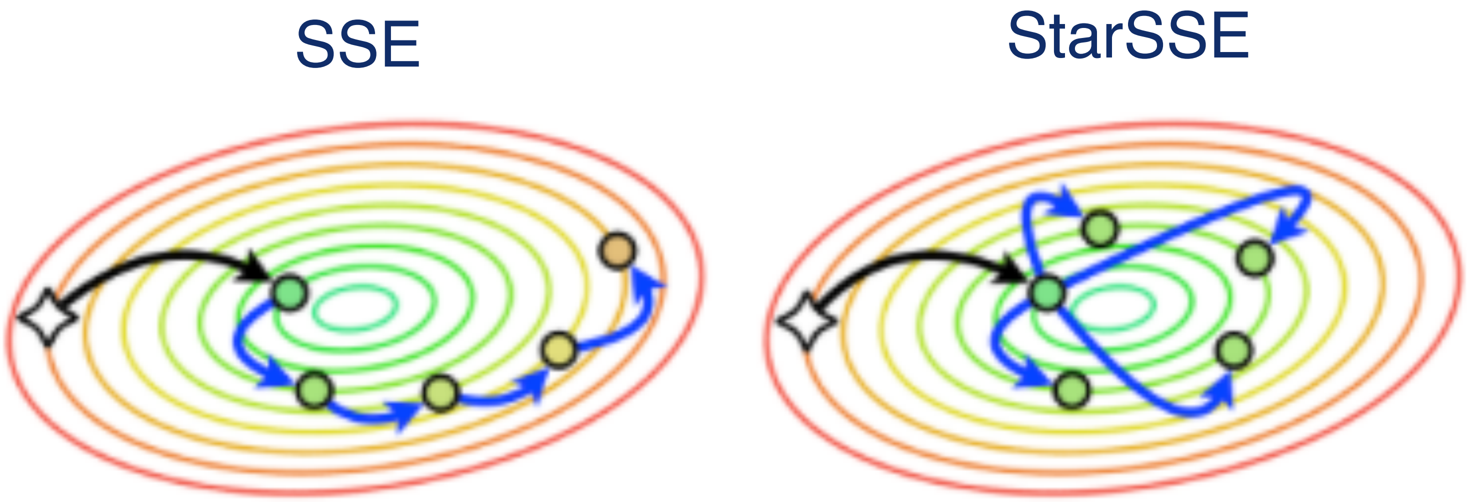
Gritsaev Timofei Grigorievich
under Ildus Sadrtdinov supervision

SNAPSHOT ENSEMBLES (SSE): TRAIN 1, GET M FOR FREE

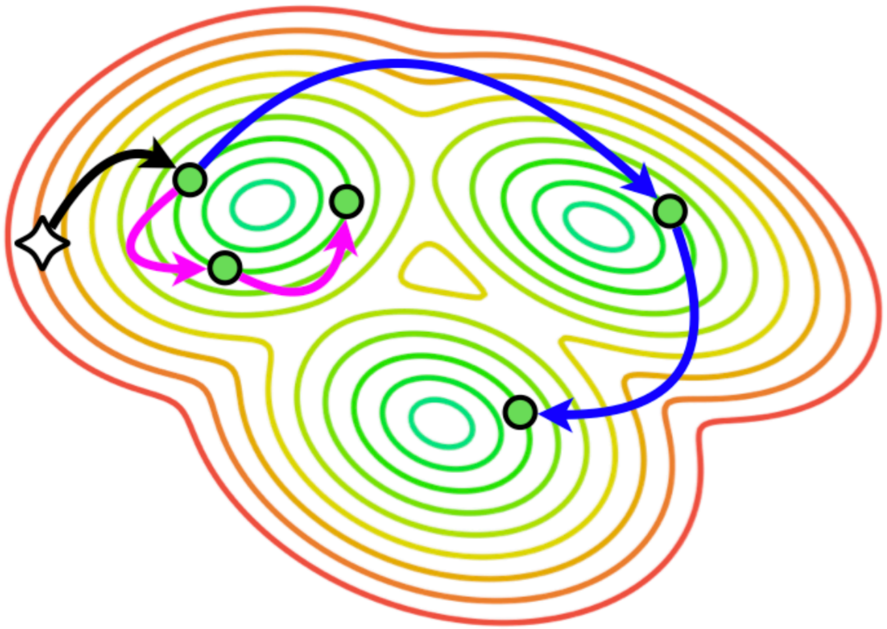




"To Stay or Not to Stay in the Pre-train Basin: Insights on Ensembling in Transfer Learning"



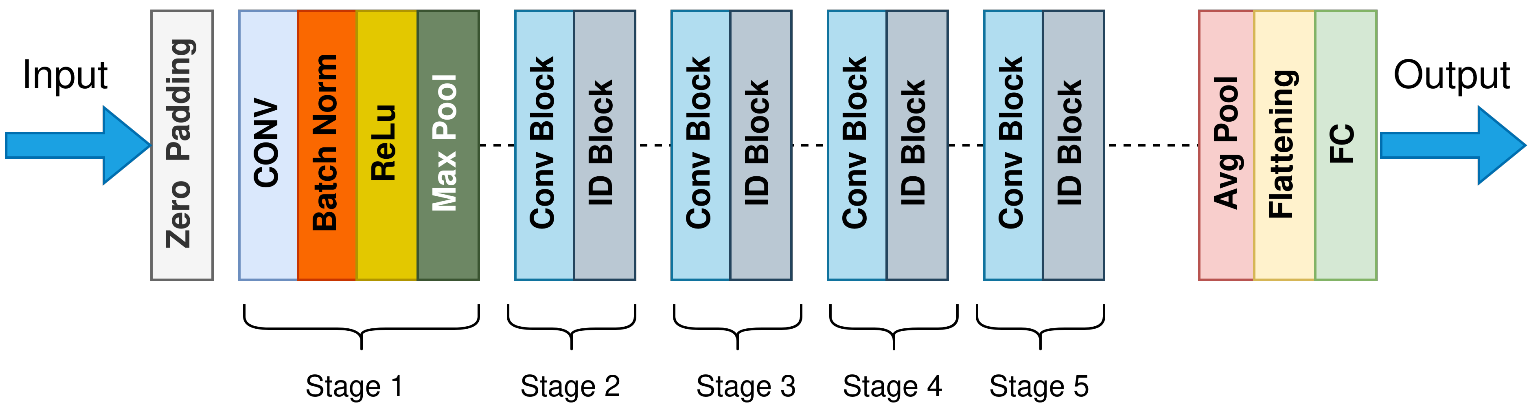
More local vs Semi-local



Methodology

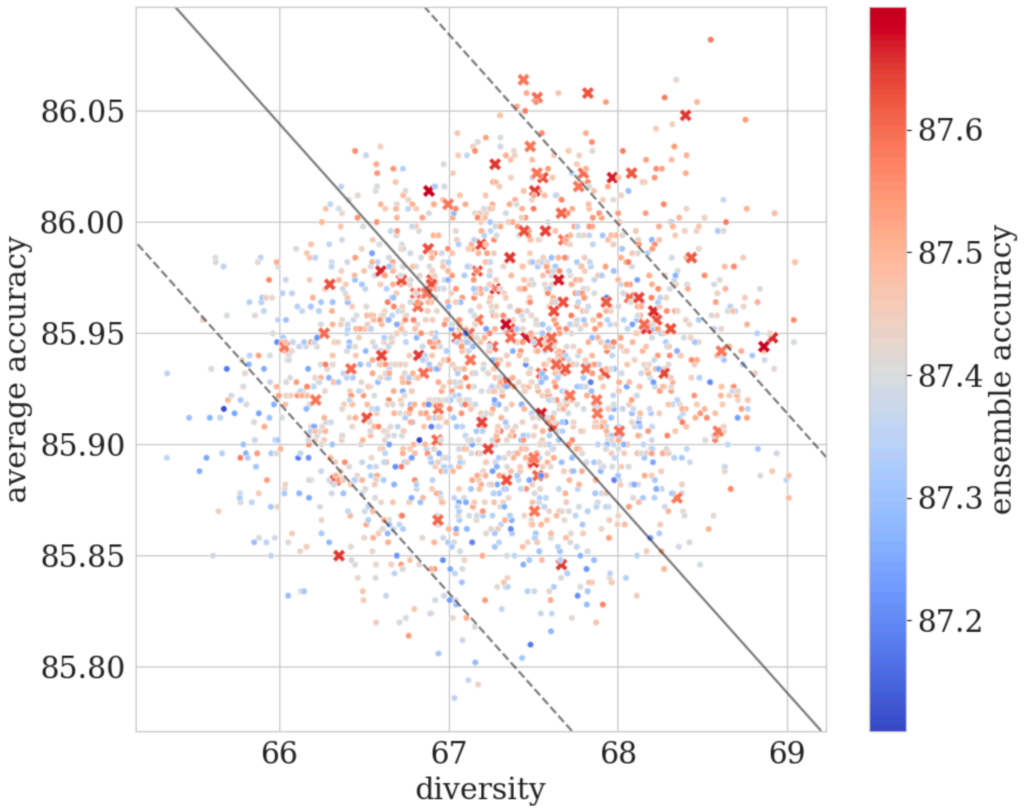
- ResNet50
- CIFAR-100
- Following the protocol from
  “To Stay or Not to Stay in the Pre-train Basin: Insights on Ensembling in Transfer Learning”

ResNet50 Model Architecture



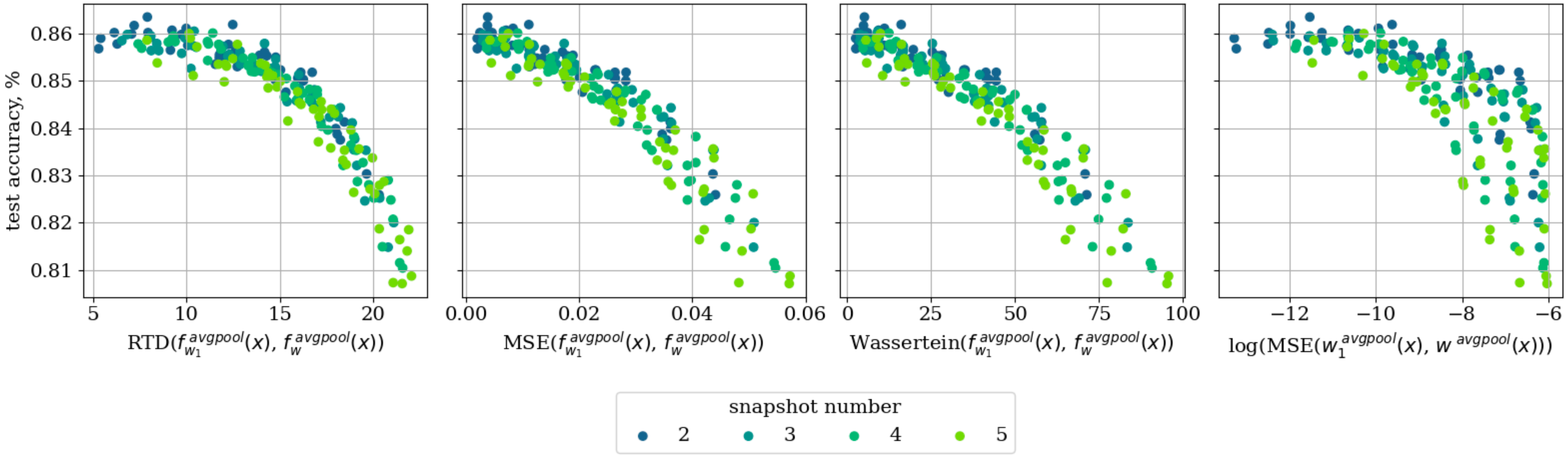
Average accuracy, diversity, ensemble accuracy

StarSSE



Algorithm		worst 5%	best 5%
StarSSE	avg.acc.	85.91±0.04	85.95±0.05
	div.	67.11±0.78	67.44±0.6
	ens. acc.	87.24±0.03	87.62±0.03

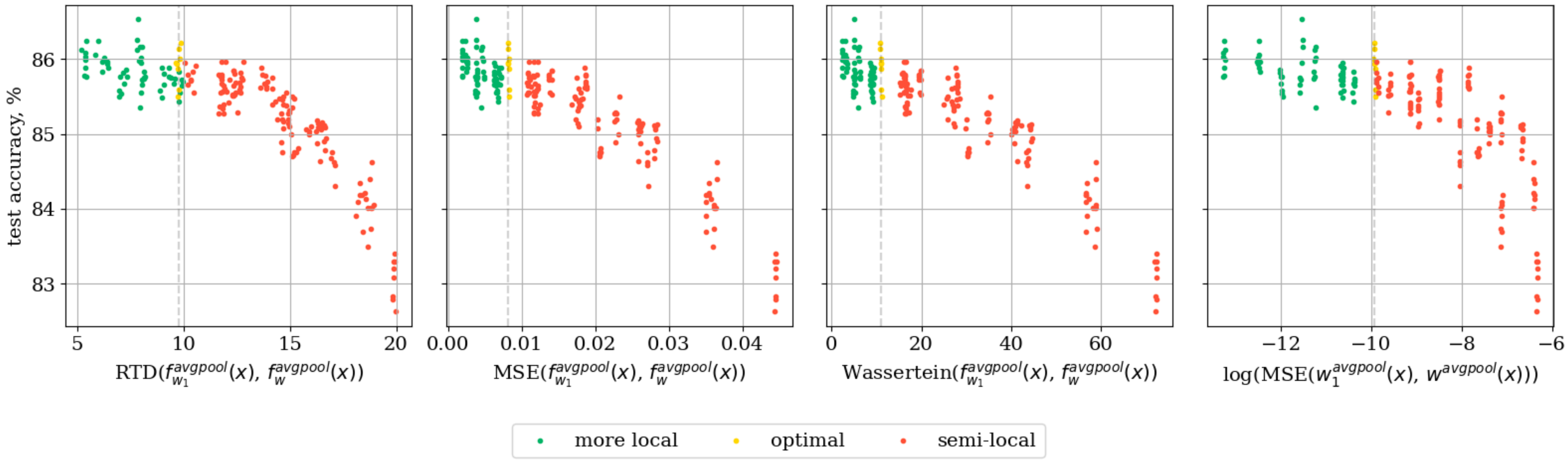
SSE regularization motivation



Regularized SSE results

Algorithm	avg. acc.	div.	**ens. acc.**
SSE	85.52±0.34	71.57±5.94	87.11±0.06
StarSSE	85.88±0.25	68.12±1.15	**87.41**±0.12
SSE–RTD	85.71±0.07	68.04±5.78	87.36±0.09
SSE–MSE	85.81±0.221	67.62±5.95	87.37±0.02

Analysis of StarSSE internal representations

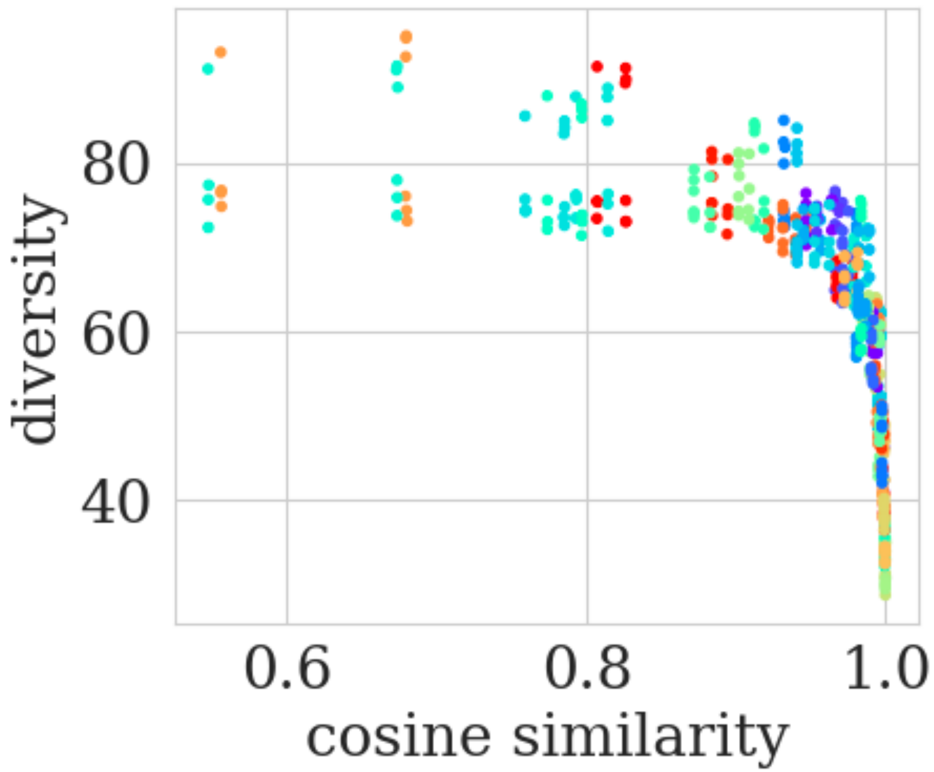


Pairwise diversity

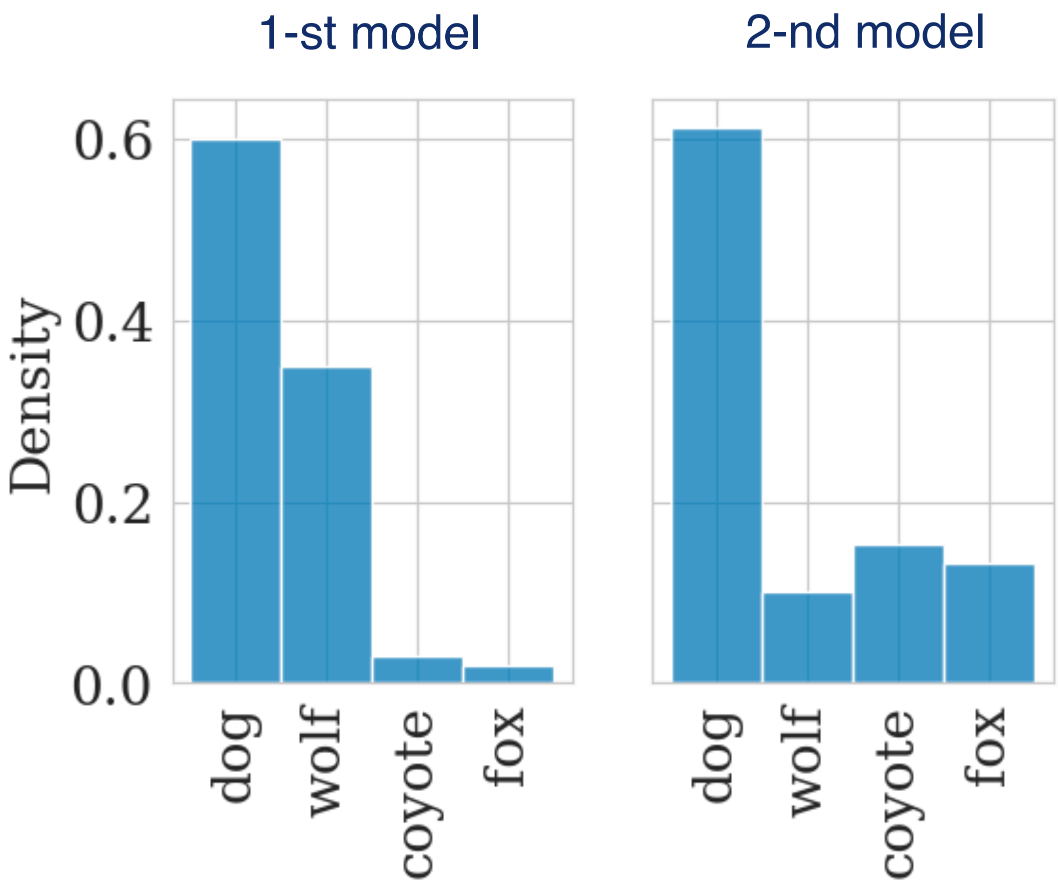
Random ensemble

	I	II	III	IV	V
I	0.0	73.26	70.83	69.58	70.76
II	73.26	0.0	70.93	72.06	**74.49**
III	70.83	70.93	0.0	71.07	71.14
IV	69.58	72.06	71.07	0.0	71.57
V	70.76	**74.49**	71.14	71.57	0.0

StarSSE—WO



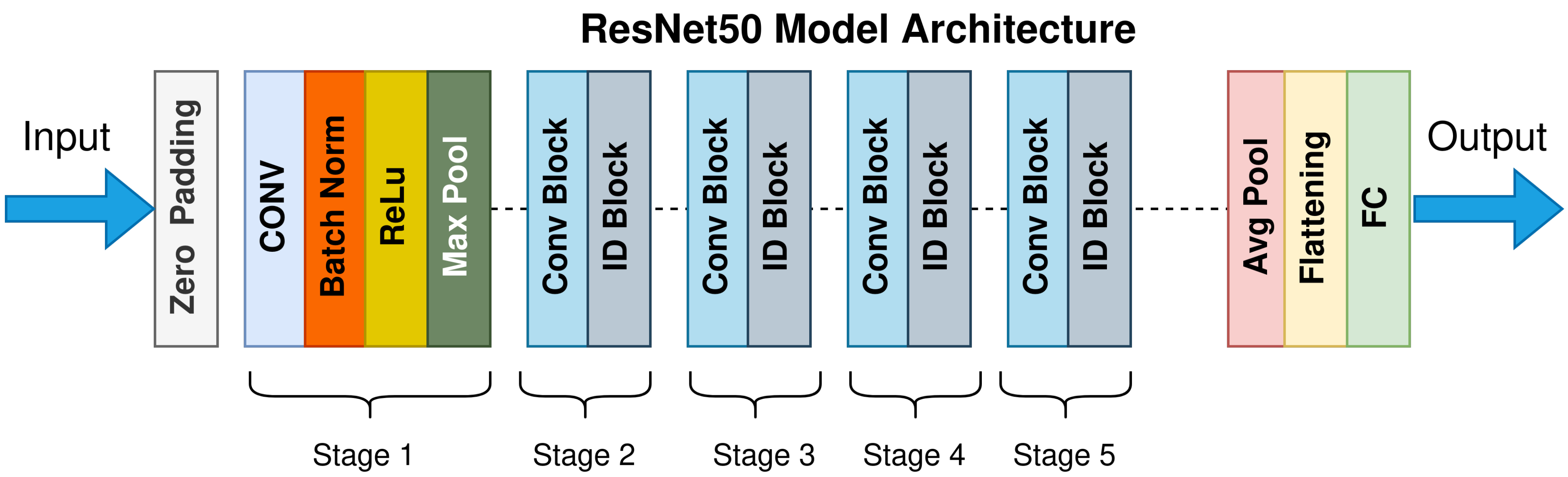
StarSSE—CE



The modifications of StarSSE results

Algorithm	avg. acc.	div.	**ens. acc.**
SSE	85.52±0.34	71.57±5.94	87.11±0.06
StarSSE	85.88±0.25	68.12±1.15	**87.41**±0.12
StarSSE–WO	85.83±0.24	67.31±1.77	87.4±0.16
StarSSE–CE (2)	85.4±0.49	75.21±3.79	**87.44**±0.2

StarSSE with Parameter efficient Fine-tuning (PeFt)



StarSSE with PeFt results

Algorithm	avg. acc.	div.	ens. acc.
StarSSE	85.88$\pm$0.25	68.12$\pm$1.15	**87.41**$\pm$0.12
StarSSE (1) x0.74 time	85.57$\pm$0.356	71.37$\pm$2.02	87.26$\pm$0.06
StarSSE (2) x0.6 time	85.94$\pm$0.225	60.94$\pm$1.25	87.1$\pm$0.14
StarSSE (3) x0.55 time	86.01$\pm$0.195	17.38$\pm$1.02	86.09$\pm$0.22

Conclusion

- Individual quality and diversity are the key
- Increasing StarSSE individual quality reduces ensemble performance
- Increasing StarSSE diversity corrupts individual models, but it is possible
- StarSSE is the best algorithm, gives the best diversity with insignificant individual quality decrease