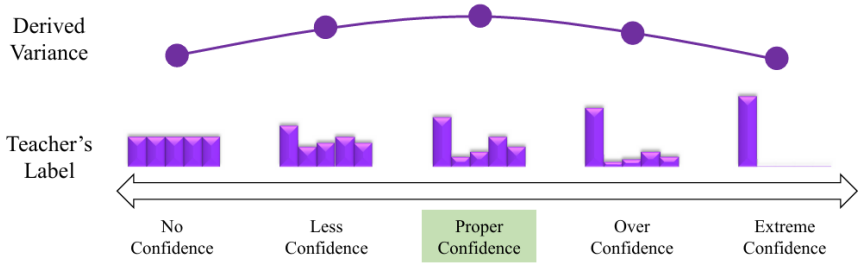
Exploring Dark Knowledge under Various Teacher Capacities and Addressing Capacity Mismatch

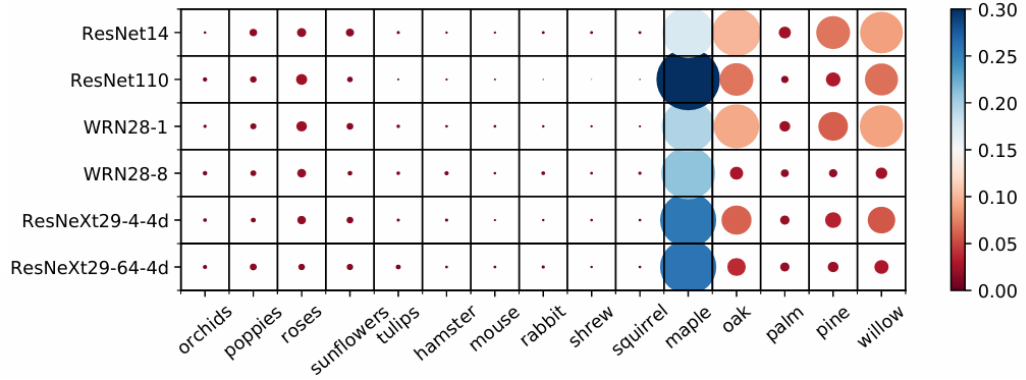
Wen-Shu FAN, Xin-Chun LI, De-Chuan ZHAN

Frontiers of Computer Science, DOI: 10.1007/s11704-025-41434-w

Problems & Ideas

- Problems of existing knowledge distillation (KD) methods:
  - Larger teacher fails to teach better student (capacity mismatch).
  - Without a fundamental grasp of dark knowledge, the causes of capacity mismatch remain elusive.
- Ideas: Two observations reflect dark knowledge: larger teachers exhibit lower variance on non-target classes, teachers with different capacities maintain relative class affinity.

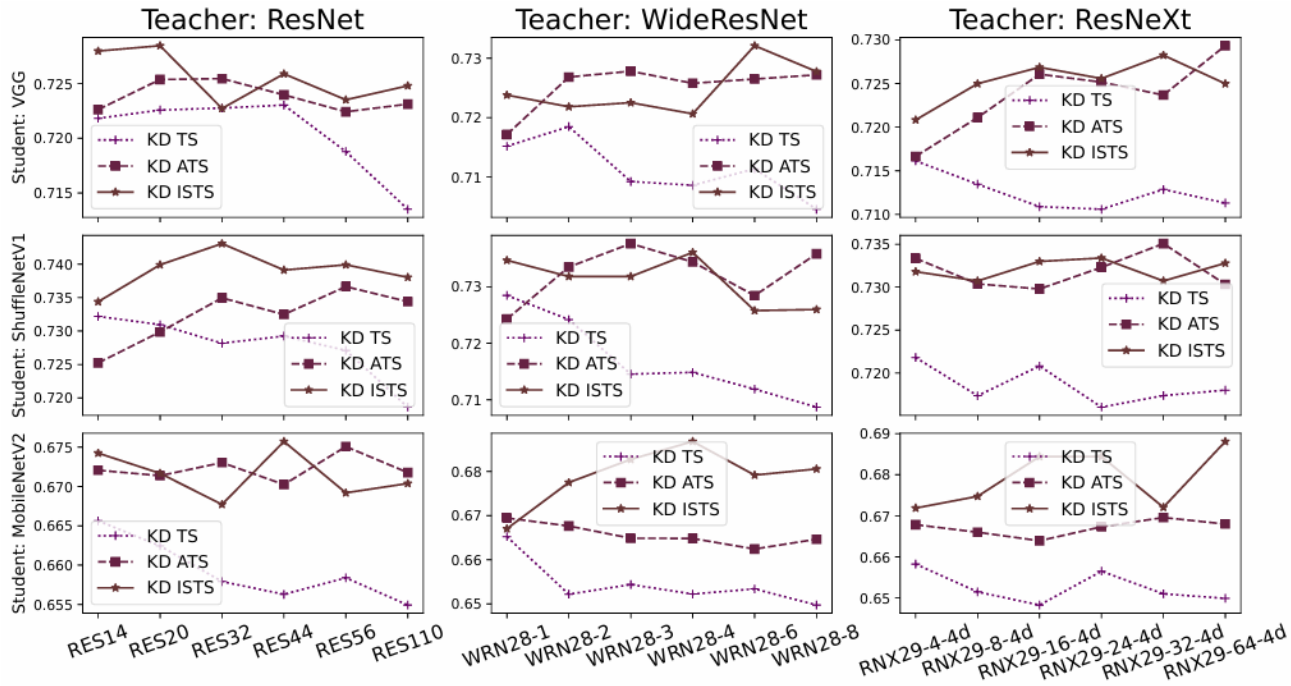




Two fundamental observations reflecting dark knowledge.
Left: When teacher becomes very large, a larger teacher tends to produce probability vectors with lower distinction among non-ground-truth classes, which neglects information among non-target classes;
Right: Teachers with different capacities are basically consistent in their cognition of relative class affinity.

Main Contributions

- Contributions:
  - We show novel insights about the dark knowledge provided by teachers with various capacities, including their distinctness on absolute class probabilities and consistency in relative class affinities;
  - We address the capacity mismatch problem in KD by proposing multiple simple yet effective methods based on the above insights, which are verified by abundant experimental studies.



Distillation results via TS, ATS, and ISATS we proposed on CIFAR-100. The x-axis of each figure shows teachers with various capacities. We can see ISATS delivers the best performance and exhibits the least pronounced capacity mismatch.