Guiding Yourself with Your Own Insights: Student-Driven Knowledge Distillation

Dacheng Qi, Huayu Zhang, Yufeng Wang, Shuangkang Fang, Zehao Zhang, Zesheng Wang, Wenrui Ding

School of Electronic and Information Engineering / Institute of Unmanned System, Beihang University

Motivation

Recent studies reveal a consistent trend: **Better architectural alignment between teacher and student consistently improves distillation performance.** This observation motivates a deeper exploration into architectural compatibility as a key factor for effective knowledge transfer.

Student			*Teacher*				*KD*
Model	ACC	CE-L	Model	ACC	τ	CE-L (τ)	ACC
VGG8	70.36	1.27	RN110	74.31	1.00	1.18114	71.51
			VGG13	74.64	2.05	1.18752	71.98
RN20	69.06	1.13	RN110	74.31	1.00	1.18114	69.80
			VGG13	74.64	2.05	1.18752	69.77

Our Contributions:

- Propose **Student-Driven Knowledge Distillation (SDKD)** using a proxy teacher to bridge teacher-student disparity.
- Design the **proxy teacher** with the same architecture as the student for better alignment of logits.
- Introduce a **Feature Fusion Block (FFB)** to enhance the proxy teacher with teacher features for richer knowledge transfer.

Student-Driven Knowledge Distillation

Deep Mutual Learning (DML) employs mutual training between teacher and student, achieving perfect architectural alignment. Mutual training allows DML to consistently surpass vanilla KD. However, involving the full teacher network is memory-intensive. We introduce a **proxy teacher** to preserve the benefits of mutual training while reducing memory cost and enhancing distillation efficiency.

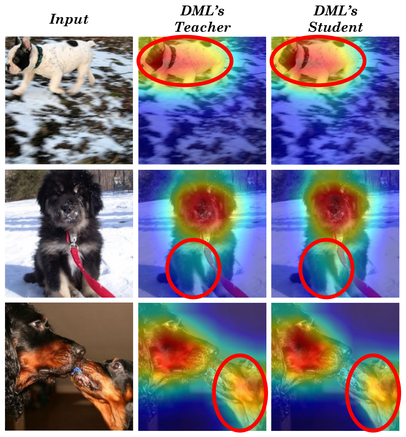
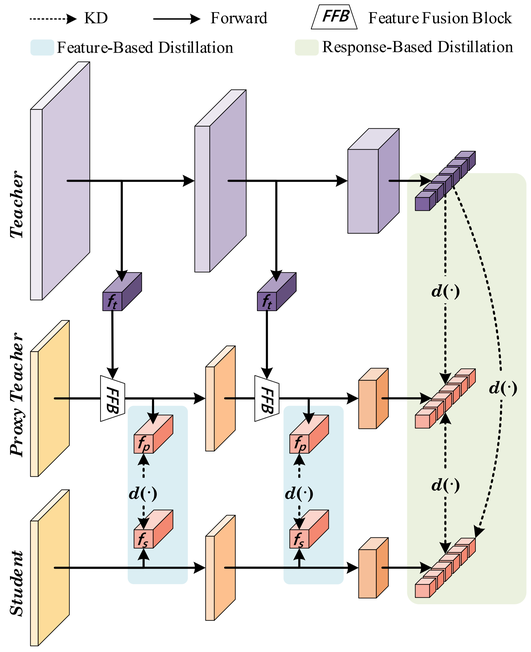
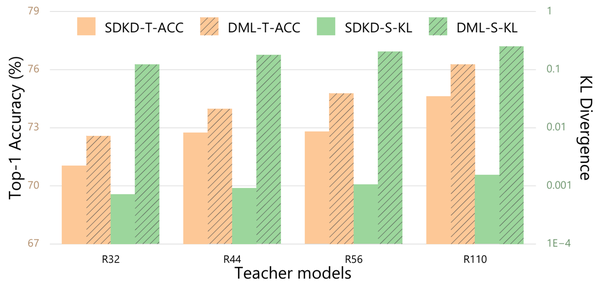


Illustration of SDKD



- **Proxy Teacher** is built by directly copying the student's architecture and weights, reducing the teacher-student gap.
- The **FFB** integrates intermediate features from the teacher and proxy to boost representation.
- Only the **FFB** is trainable, making the framework **lightweight**, **efficient**, and effective for distillation.



- **KL Divergence** quantifies the difference between the (proxy) teacher and student outputs.

Experiments

Top-1 Accuracy (%) on CIFAR-100

Teacher	*WRN-40-2*	*WRN-40-2*	*ResNet56*	*ResNet110*	*ResNet110*
Student	*WRN-16-2*	*WRN-40-1*	*ResNet20*	*ResNet20*	*ResNet32*
Teacher	75.61	75.61	72.34	74.31	74.31
Student	73.26	71.98	69.06	69.06	71.14
KD	74.92	73.54	70.66	70.67	73.08
DML	75.33	74.24	71.48	71.52	73.59
FitNet	73.58	72.24	69.21	68.99	71.06
AT	74.08	72.77	70.55	70.22	72.31
SP	73.83	72.43	69.67	70.04	72.69
CC	73.56	72.21	69.63	69.48	71.48
VID	74.11	73.30	70.38	70.16	72.61
RKD	73.35	72.22	69.61	69.25	71.82
PKT	74.54	73.45	70.34	70.25	72.61
AB	72.50	72.38	69.47	69.53	70.98
FT	73.25	71.59	69.84	70.22	72.37
FSP	72.91	-	69.95	70.11	71.89
NST	73.68	72.24	69.60	69.53	71.96
CRD	75.48	74.14	71.16	71.46	73.48
SRRL	75.49	74.64	70.86	70.78	73.21
SemCKD	-	74.41	-	-	-
SimKD	-	**75.56**	-	-	-
DistPro	**76.36**	-	72.03	-	73.74
NORM	75.65	74.82	71.35	71.55	73.67
SoTeacher	75.39	74.35	71.32	71.27	73.77
SDKD(ours)	75.85	75.00	**72.11**	**72.13**	**74.24**
Gain	2.59	3.02	3.05	3.07	3.10

Visualization of attention regions