Good Teachers Explain:
Explanation-enhanced Knowledge Distillation

Amin Parchami-Araghi*, Moritz Böhle*, Sukrut Rao*, Bernt Schiele

MAX PLANCK INSTITUTE FOR INFORMATICS

SIC Saarland Informatics Campus

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION

Summary

Knowledge Distillation (KD) is effective for student accuracy.

But

- teacher-student **agreement can be low**.
- teacher's **reasoning might not be distilled**.
- teacher's **interpretability might get lost**.

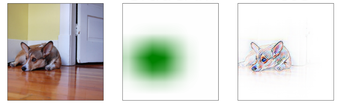
Agreement := T: ✓ ✓ ✓ ✓ ✓ ✗ ✗ ✓ ✗
S: ✗ ✓ ✗ ✗ ✓ ✓ ✓ ✓ ✓

By simply matching explanations (e²KD)

- teacher-student **agreement and accuracy improve**.
- students **remain *right for right reasons.***
- students **remain interpretable.**

Experimental Setup

Explanation Methods



Teacher→Student

RN34 →RN18
RN50 → [RN18, ConvNext, EfficientNet, MobileNet]
B-cos RN34 →B-cos RN18
B-cos DN169 → [B-cos ViT$_{\text{Tiny}}$, B-cos RN18]

Datasets ImageNet, Waterbirds, Pascal VOC, SUN397

e²KD: Explanation-enhanced KD

In addition to logits

$$\mathcal{L}_{KD} = -\tau^2 \sum_{j=1}^{C} \sigma_j\left(\frac{z^T}{\tau}\right) \log \sigma_j\left(\frac{z^S}{\tau}\right)$$

also match explanations:

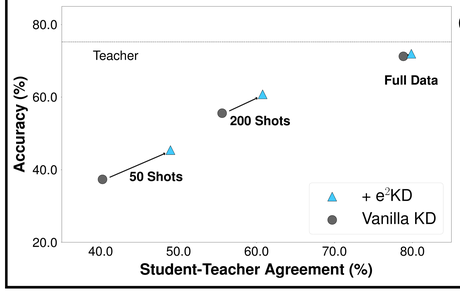
$$\mathcal{L}_{exp} = 1 - \cos\left(\mathbf{Explain}(T, x, \hat{y}_T), \mathbf{Explain}(S, x, \hat{y}_T)\right)$$

✅ **Leverages existing explanation methods**

✅ **Model-agnostic** ✅ **Parameter-free**

We discuss three desiderata for faithful distillation

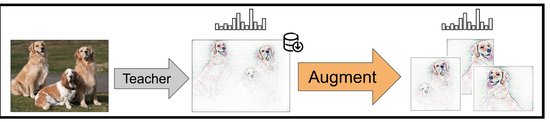
High Agreement with the Teacher



Consistent acc. and agreement gains
+ especially on limited-data
+ even on unrelated images

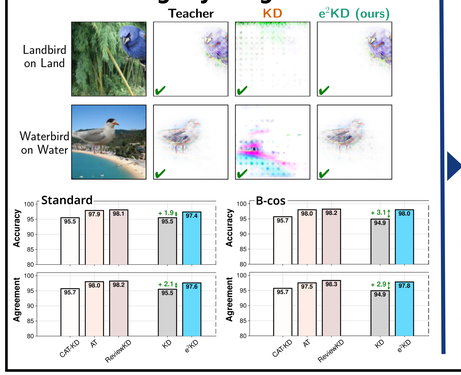
IMN → SUN	Acc.	Agr.
Teacher	60.5	-
Using SUN	57.7	67.9
KD	53.5	65.0
+ e²KD	**54.9**	**67.7**

Bonus: ❄ e²KD

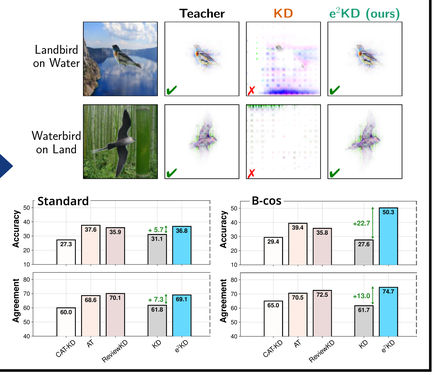


Similar Predictions, but for Similar Reasons!

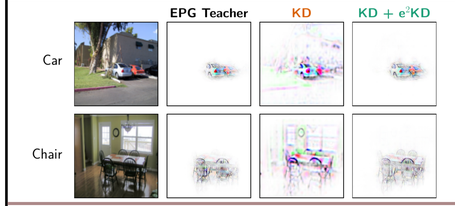
Remain *right for right reasons*



Remain OOD-Robust

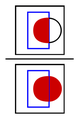


Maintain Teacher's Interpretability

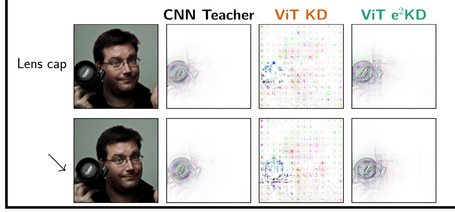


Learn to localize (VOC Dataset)

EPG Score



	EPG	IoU	F1
Teacher	75.7	21.3	72.5
Baseline	50.0	29.0	58.0
KD	60.1	31.6	60.1
+ e²KD	**71.1**	24.8	**67.6**



Learn inductive biases (CNN →ViT)

Method	Acc.	Agr.
T: B-cos DenseNet-169	75.2	-
B: B-cos ViT$_{\text{Tiny}}$	60.0	64.6
KD	64.8	70.1
+ e²KD	**66.3**	**71.8**

References Agreement *(Stanton et al., 2021)*; B-cos *(Böhle et al., 2021)*; GradCAM *(Selvaraju et al., 2017)*; RRR (*Ross et. al 2017.*); Guided Teachers *(Rao et al., 2023)*; EPG *(Wang et al., 2020)*; Waterbirds *(Sagawa et al., 2019)*