Good Teachers Explain: Explanation-Enhanced Knowledge Distillation

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

Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken
{mparcham,mboehle,sukrut.rao,schiele}@mpi-inf.mpg.de

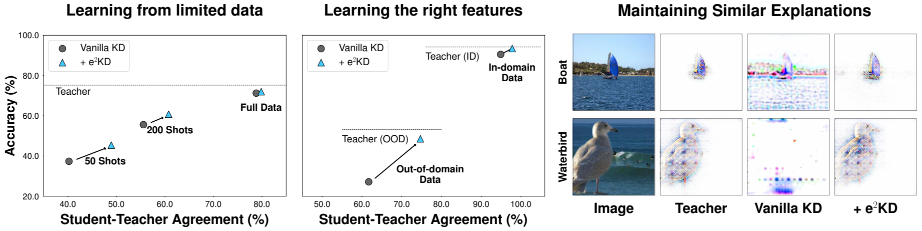


Fig. 1: A good teacher explains. Using explanation-enhanced KD (**e^{2}KD**) improves distillation faithfulness and student performance. E.g., e^{2}KD allows the student models to more faithfully approximate the teacher, especially when using fewer data, leading to large gains in accuracy and teacher-student agreement **(left)**. Further, by guiding the students to give similar explanations as the teacher, e^{2}KD ensures that students learn to be 'right for the right reasons', improving their accuracy under distribution shifts **(center)**. Lastly, e^{2}KD students learn similar explanations as the teachers, thus exhibiting a similar degree of interpretability as the teacher **(right)**.

Abstract. Knowledge Distillation (KD) has proven effective for compressing large teacher models into smaller student models. While it is well known that student models can achieve similar accuracies as the teachers, it has also been shown that they nonetheless often do not learn the same function. It is, however, often highly desirable that the student's and teacher's functions share similar properties such as basing the prediction on the same input features, as this ensures that students learn the 'right features' from the teachers. In this work, we explore whether this can be achieved by not only optimizing the classic KD loss but also the similarity of the explanations generated by the teacher and the student. Despite the idea being simple and intuitive, we find that our proposed 'explanation-enhanced' KD (e^{2}KD) (1) consistently provides large gains over logit-based KD in terms of accuracy and student-teacher agreement, (2) ensures that the student learns from the teacher to be right for the right reasons and to give similar explanations, and (3) is robust with respect to the model architectures, the amount of training data, and even works with 'approximate', pre-computed explanations.

Keywords: Model Compression · Faithful Distillation · Interpretability

[*] Denotes equal contribution. Code: github.com/m-parchami/GoodTeachersExplain

1 Introduction

Knowledge Distillation (KD) [17] has proven effective for improving classification accuracies of relatively small 'student' models, by training them to match the logit distribution of larger, more powerful 'teacher' models. Despite its simplicity, this approach can be sufficient for the students to match the teacher's accuracy, while requiring only a fraction of the computational resources of the teacher [4]. Recent findings, however, show that while the students might match the teacher's accuracy, the knowledge is nonetheless not distilled faithfully [32].

Faithful KD, i.e. a distillation that ensures that the teacher's and the student's functions share properties beyond classification accuracy, is however desirable for many reasons. E.g., the lack of model agreement [32] can hurt the user experience when updating machine-learning-based applications [3,36]. Similarly, if the students use different input features than the teachers, they might not be *right for the right reasons* [27]. Further, given the recent AI Act proposal by European legislators [9], it is likely that model interpretability will play an increasingly important role and become an intrinsic part of the model functionality. To maintain the *full* functionality of a model, KD should thus ensure that the students allow for the same degree of model interpretability as the teachers.

To address this, in this work we discuss three desiderata for faithful KD and study if promoting explanation similarity using commonly used model explanations such as GradCAM [29] or those of the recently proposed B-cos models [5] can increase the faithfulness of distillation. This should be the case if such explanations indeed reflect meaningful aspects of the models' 'internal reasoning'. Concretely, we propose **'explanation-enhanced' KD (e^{2}KD)**, a simple, parameter-free, and model-agnostic addition to KD in which we train the student to also match the teacher's explanations.

Despite its simplicity, e^{2}KD significantly advances towards faithful distillation in a variety of settings (Fig. 1). Specifically, e^{2}KD improves student accuracy, ensures that the students learn to be right for the right reasons, and inherently promotes consistent explanations between teachers and students. Moreover, the benefits of e^{2}KD are robust to limited data, approximate explanations, and across model architectures. In short, we make the **following contributions**:
(1) We propose **explanation-enhanced KD (e^{2}KD)** and train the students to not only match the teachers' logits, but also their explanations (Sec. 3.1); for this, we use B-cos and GradCAM explanations. This not only yields competitive students in terms of accuracy, but also significantly improves KD faithfulness on the ImageNet [10], Waterbirds-100 [22,28], and PASCAL VOC [11] datasets.
(2) We discuss three desiderata for measuring the faithfulness of KD (Sec. 3.2). We evaluate whether the student is performant and has high agreement with the teacher (Desideratum 1), examine whether students learn to use the same input features as a teacher that was guided to be 'right for the right reasons' even when distilling with biased data (Desideratum 2), and explore whether they learn the same explanations and architectural priors as the teacher (Desideratum 3).
(3) We show e^{2}KD to be a robust approach for improving knowledge distillation, which provides consistent gains across model architectures and with limited

data. Further, e^{2}KD is even robust to using cheaper 'approximate' explanations. Specifically, for this we propose '**frozen explanations**' which are only computed once and, during training, undergo the same augmentations as images (Sec. 3.3).

2 Related Work

Knowledge Distillation (KD) has been introduced to compress larger models into more efficient models for cost-effective deployment [17]. Various approaches have since been proposed, which we group into three types in the following discussion: *logit-* [4,17,42], *feature-* [8,26,31,39], and *explanation-based KD* [1,15,40].

Logit-based KD [17], which optimizes the logit distributions of teacher and student to be similar, can suffice to match their accuracies, as long as the models are trained for long enough ('patient teaching') and the models' logits are based on the same images ('consistent teaching'), see [4]. However, [32] showed that despite such a careful setup, the function learnt by the student can still significantly differ from the teacher's by comparing the agreement between the two. We expand on [32] and introduce additional settings to assess the faithfulness of distillation, and show that it can be significantly improved by a surprisingly simple explanation-matching approach. While [21] finds that KD does seem to transfer additional properties to the student, by showing that GradCAM explanations of the students are more similar to the teacher's than those of an independently trained model, we show that explicitly optimizing for explanation similarity significantly improves this w.r.t. logit-based KD, whilst also yielding important additional benefits such as higher robustness to distribution shifts.

Feature-based KD approaches [8,19,26,31,39] provide additional information to the students by optimizing some of the students' intermediate activation maps to be similar to those of the teacher. For this, specific choices regarding which layers of teachers and students to match need to be made and these approaches are thus architecture-dependent. In contrast, our proposed e^{2}KD is architecture-agnostic as it matches only the explanations of the models' predictions.

Explanation-based KD approaches have only recently begun to emerge [1,15,40] and these are conceptually most related to our work. In CAT-KD [15], the authors match class activation maps (CAM [43]) of students and teachers. As such, CAT-KD can also be considered an 'explanation-enhanced' KD (e^{2}KD) approach. However, the explanation aspect of the CAMs plays only a secondary role in [15], as the authors even reduce the resolution of the CAMs to 2×2 and faithfulness is not considered. In contrast, we explicitly introduce e^{2}KD to promote faithful distillation and evaluate faithfulness across multiple settings. Further, similar to our work, [1] argues that explanations can form part of the model functionality and should be considered in KD. For this, the authors train an additional autoencoder to mimic the explanations of the teacher; explanations and predictions are thus produced by separate models. In contrast, we optimize the students directly to yield similar explanations as the teachers in a simple and parameter-free manner.

Fixed Teaching. [12, 30, 38] explore pre-computing the logits at the start of training to limit the computational costs due to the teacher. In addition to pre-computing *logits*, we pre-compute *explanations* and show how they can nonetheless be used to guide the student model during distillation.

Explanation Methods. To better understand the decision making process of DNNs, many explanation methods have been proposed in recent years [2, 5, 25, 29]. For our e^{2}KD experiments, we take advantage of the differentiability of attribution-based explanations and train the student models to yield similar explanations as the teachers. In particular, we evaluate both a popular post-hoc explanation method (GradCAM [29]) as well as the model-inherent explanations of the recently proposed B-cos models [5, 6].

Model Guidance. e^{2}KD is inspired by recent advances in model guidance [13, 14, 22, 24, 27], where models are guided to focus on desired input features via human annotations. Analogously, we also guide the focus of student models, but using knowledge (explanations) of a teacher model instead of a human annotator. As such, no explicit guidance annotations are required in our approach. Further, in contrast to the discrete annotations typically used in model guidance (e.g. bounding boxes or segmentation masks), we use the real-valued explanations as given by the teacher model. Our approach thus shares additional similarities with [22], in which a model is guided via the attention maps of a vision-language model. Similar to our work, the authors show that this can guide the students to focus on the 'right' input features. We extend such guidance to KD and discuss the benefits that this yields for faithful distillation.

3 Explanation-Enhanced KD and Evaluating Faithfulness

To promote faithful KD, we introduce our proposed *explanation-enhanced KD* (e^{2}KD) in Sec. 3.1. Then, in Sec. 3.2, we present three desiderata that faithful KD should fulfill and why we expect e^{2}KD to be beneficial in the presented settings. Finally, in Sec. 3.3, we describe how to take advantage of e^{2}KD even without querying the teacher more than once per image when training the student.

Notation. For model M and input x, we denote the predicted class probabilities by $p_M(x)$, obtained using softmax $\sigma(.)$ over output logits $z_M(x)$, possibly scaled by temperature τ. We denote the class with highest probability by $\hat{y}_M$.

3.1 Explanation-Enhanced Knowledge Distillation

The logit-based knowledge distillation loss $\mathcal{L}_{KD}$ which minimizes KL-Divergence D_{KL} between teacher T and student S output probabilities is given by

$$\mathcal{L}_{KD} = \tau^2 D_{\mathrm{KL}}(p_T(x;\tau)||p_S(x;\tau)) = -\tau^2 \sum_{j=1}^{c} \sigma_j\left(\frac{z_T}{\tau}\right) \log \sigma_j\left(\frac{z_S}{\tau}\right). \tag{1}$$

We propose to leverage advances in model explanations and explicitly include a term $\mathcal{L}_{exp}$ that promotes explanation similarity for a more faithful distillation:

$$\mathcal{L} = \mathcal{L}_{KD} + \lambda\mathcal{L}_{exp}. \tag{2}$$

Specifically, we maximize the similarity between the models' explanations, for the class $\hat{y}_T$ predicted by the teacher:

$$\mathcal{L}_{exp} = 1 - \text{sim}\left(E(T, x, \hat{y}_T), E(S, x, \hat{y}_T)\right) \ . \tag{3}$$

Here, $E(M, x, \hat{y}_T)$ denotes an explanation of model M for class $\hat{y}_T$ and sim a similarity function; in particular, we rely on well-established explanation methods (e.g. GradCAM [29]) and use cosine similarity in our experiments.

e^{2}KD is model-agnostic. Note that by computing the loss only across model outputs and explanations, e^{2}KD does not make any reference to architecture-specific details. In contrast to feature distillation approaches, which match specific blocks between teacher and student, e^{2}KD thus holds the potential to seamlessly work across different architectures without any need for adaptation. As we show in Sec. 4, this indeed seems to be the case, with e^{2}KD improving the distillation faithfulness out of the box for a variety of model architectures, such as CNNs, B-cos CNNs, and even B-cos ViTs [6].

3.2 Evaluating Benefits of e^{2}KD

In this section, we discuss three desiderata that faithful KD should fulfill and why we expect e^{2}KD to be beneficial. While distillation methods are often compared in terms of accuracy, our findings (Sec. 4) suggest that one should also consider the following desiderata to judge a distillation method on its faithfulness.

Desideratum 1: **High Agreement with Teacher.** First and foremost, faithful KD should ensure that the student classifies any given sample in the same way as the teacher, i.e., the student should have high agreement [32] with the teacher. For inputs $\{x_i\}_{i=1}^N$ this is defined as:

$$\text{Agreement}(T, S) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\hat{y}_{i,T}=\hat{y}_{i,S}} \ . \tag{4}$$

While [32] found that more data points can improve the agreement, in practice, the original dataset that was used to train the teacher might be proprietary or prohibitively large (e.g. [23]). It can thus be desirable to effectively distill knowledge *efficiently* with less data. To assess the effectiveness of a given KD approach in such a setting, we propose to use a teacher trained on a large dataset (e.g. ImageNet [10]) and distill its knowledge to a student using as few as 50 images per class (≈ 4% of the data) or even on images of an unrelated dataset.

Compared to standard supervised training, it has been argued that KD improves the student performance by providing more information (full logit distribution instead of binary labels). Similarly, by additionally providing the teachers' explanations, we show that e^{2}KD boosts the performance even further, especially when fewer data is available to learn the same function as the teacher (Sec. 4.1).

Desideratum 2: Learning the 'Right' Features. Despite achieving high accuracy, models often rely on spurious input features (are not "right for the right reasons" [27]), and can generalize better if guided to use the 'right' features via human annotations. This is particularly useful in the presence of distribution shifts [28]. Hence, faithful distillation should ensure that student models also learn to use these 'right' features from a teacher that uses them.

To assess this, we use a binary classification dataset [28] in which the background is highly correlated with the class label in the training set, making it challenging for models to learn to use the actual class features for classification. We use a teacher that has explicitly been guided to focus on the actual class features and to ignore the background. Then, we evaluate the student's accuracy and agreement with the teacher under distribution shift, i.e., at test time, we evaluate on images in which the class-background correlation is reversed. By providing additional spatial clues from the teachers' explanations to the students, we find that e^{2}KD significantly improves performance over KD (Sec. 4.2).

Desideratum 3: Maintaining Interpretability. Note that the teachers might be trained to exhibit certain desirable properties in their explanations [24], or do so as a result of a particular training paradigm [7] or the model architecture [6].

We propose two settings to test if such properties are transferred. First, we measure how well the students' explanations reflect properties the teachers were explicitly trained for, i.e. how well they localize class-specific input features when using a teacher that has explicitly been guided to do so [24]. We find e^{2}KD to lend itself well to maintaining the interpretability of the teacher, as the explanations of students are explicitly optimized for this (Sec. 4.3 'Distill on VOC').

Secondly, we perform a case study to assess whether KD can transfer priors that are not *learnt*, but rather inherent to the model architecture. Specifically, the explanations of B-cos ViTs have been shown to be sensitive to image shifts [6], even when shifting by just a few pixels. To mitigate this, the authors of [6] proposed to use a short convolutional stem. Interestingly, in Sec. 4.3 'Distill to ViT', we find that by learning from a CNN teacher under e^{2}KD, the explanations of a ViT student without convolutions also become largely equivariant to image shifts, and exhibit similar patterns as the teacher.

3.3 e^{2}KD with 'Frozen' Explanations

Especially in the 'consistent teaching' setup of [4], KD requires querying the teacher for every training step, as the input images are repeatedly augmented. To reduce the computational cost incurred by evaluating the teacher, recent work explores using a 'fixed teacher' [12, 30, 38], where logits are pre-computed once at the start of training and used for all augmentations.

Analogously, we propose to use pre-computed explanations for images in the e^{2}KD framework. For this, we apply the same augmentations (e.g. cropping or flipping) to images and the teacher's explanations during distillation. In Sec. 4.4, we show that e^{2}KD is robust to such 'frozen' explanations, despite the fact that they of course only approximate the teacher's explanations. As such, frozen explanations provide a trade-off between optimizing for explanation similarity and reducing the cost due to the teacher.

4 Results

In the following, we present our results. Specifically, in Sec. 4.1 we compare KD approaches in terms of accuracy and agreement on ImageNet as a function of the distillation dataset size. Thereafter, we present the results on learning the 'right' features from biased data in Sec. 4.2 and on maintaining the interpretability of the teacher models in Sec. 4.3. Lastly, in Sec. 4.4, we show that e^{2}KD can also yield significant benefits with approximate 'frozen' explanations (cf. Sec. 3.3).

Before turning to the results, however, we first provide some general details with respect to explanation methods used for e^{2}KD and our training setup.

Explanation methods. For e^{2}KD, we use GradCAM [29] for standard models and B-cos explanations for B-cos models, optimizing the cosine similarity as per Eq. (3). For B-cos, we use the dynamic weights $\mathbf{W}(\mathbf{x})$ as explanations [5].

Training details. In general, we follow the recent KD setup from [4], which has shown significant improvements for KD; results based on the setup followed by [8, 15, 39] can be found in the supplement. Unless specified otherwise, we use the AdamW optimizer [20] and, following [5], do not use weight decay for B-cos models. We use a cosine learning rate schedule with initial warmup for 5 epochs. For the teacher-student logit loss on multi-label VOC dataset, we use the logit loss following [37] instead of Eq. (1). For AT [39], CAT-KD [15], ReviewKD [8], and CRD [33] we follow the original implementation and use cross-entropy based on the ground truth labels instead of Eq. (1); for an adaptation to B-cos models, see appendix C.2. For each method and setting, we report the results of the best hyperparameters (softmax temperature and the methods' loss coefficients) as obtained on a separate validation set. Unless specified otherwise, we augment images via random horizontal flipping and random cropping with a final resize to 224 × 224. For full details, see appendix C.1.

4.1 e^{2}KD Improves Learning from Limited Data

Setup. To test the robustness of e^{2}KD with respect to the dataset size (Sec. 3.2, Desideratum 1), we distill with 50 ($\approx 4\%$) or 200 ($\approx 16\%$) shots per class, and the full ImageNet training data; further, we also distill without access to ImageNet, performing KD on SUN397 [35], whilst still evaluating on ImageNet (and vice versa). We distill ResNet-34 [16] teachers to ResNet-18 students for standard and B-cos models (Tabs. 1 and 2); additionally, we use a B-cos DenseNet-169 [18] teacher (Tab. 3) to evaluate distillation across architectures. For reference, we also provide results we obtained via AT [39], CAT-KD [15], and ReviewKD [8].

Results. In Tabs. 1 to 3, we show that e^{2}KD can significantly improve logit-based KD in terms of top-1 accuracy as well as top-1 teacher-agreement on ImageNet. We observe particularly large gains for small distillation dataset sizes.

Table 1: KD on ImageNet for standard models. For a ResNet-34 teacher and a ResNet-18 student, we show the accuracy and agreement of various KD approaches for three different distillation dataset sizes. Across all settings, e^{2}KD yields significant accuracy and agreement gains over logit-based KD approaches (KD [4, 17] and CRD [33]). Similar results are also observed for B-cos models, see Tabs. 2 and 3.

Standard Models	**50 Shots**		**200 Shots**		**Full data**	
Teacher ResNet-34 Accuracy 73.3%	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.
Baseline ResNet-18	23.3	24.8	47.0	50.2	69.8	76.8
AT [39]	38.3	41.1	54.7	59.0	69.7	74.9
ReviewKD [8]	51.2	55.6	63.0	69.0	71.4	80.0
CAT-KD [15]	32.2	34.5	55.7	60.7	70.9	78.7
KD [4, 17]	49.8	55.5	63.1	71.9	**71.8**	81.2
+ **e^{2}KD (GradCAM)**	**54.9**	**61.7**	**64.1**	**73.2**	**71.8**	**81.6**
	+5.1	**+6.2**	**+1.0**	**+1.3**	**+0.0**	**+0.4**
CRD [33]	30.0	31.8	51.0	54.9	69.4	74.6
+ **e^{2}KD (GradCAM)**	**34.7**	**37.1**	**54.1**	**58.7**	**70.5**	**76.5**
	+4.7	**+5.3**	**+3.1**	**+3.8**	**+1.1**	**+1.9**

Table 2: KD on ImageNet for B-cos models. For a B-cos ResNet-34 teacher and a B-cos ResNet-18 student, we show the accuracy and agreement of KD approaches for three different distillation dataset sizes. Across all settings, e^{2}KD significantly improves accuracy and agreement over vanilla KD, whilst remaining competitive with prior work.

B-cos Models	**50 Shots**		**200 Shots**		**Full data**	
Teacher ResNet-34 Accuracy 72.3%	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.
Baseline ResNet-18	32.6	35.1	53.9	59.4	68.7	76.9
AT [39]	41.9	45.6	57.2	63.7	69.0	77.2
ReviewKD [8]	47.5	53.2	54.1	60.8	57.0	64.6
CAT-KD [15]	53.1	59.8	58.6	66.4	63.9	73.7
KD [4, 17]	35.3	38.4	56.5	62.9	70.3	79.9
+ **e^{2}KD (B-cos)**	**43.9**	**48.4**	**58.8**	**66.0**	**70.6**	**80.3**
	+8.6	**+10.0**	**+2.3**	**+3.1**	**+0.3**	**+0.4**

E.g., for KD, accuracy and agreement for conventional (and B-cos) models on 50 shots improve by 5.1 (B-cos: 8.6) and 6.2 (B-cos: 10.0) p.p. respectively. As e^{2}KD is model-agnostic, we found consistent trends with another teacher (cf. Tab. 3), and further find it to generalise also to other distillation methods (Tab. 1; CRD).

In Tab. 5 (right), we show that e^{2}KD also provides significant gains when using unrelated data [4], improving student's ImageNet accuracy and agreement by 4.9 and 5.4 p.p. respectively, despite computing the explanations on images of SUN [35] dataset (i.e. SUN→ImageNet). Similar gains can be observed when using ImageNet images to distill a teacher trained on SUN (i.e. ImageNet→SUN).

Table 3: KD and 'frozen' KD (❄) on ImageNet for B-cos models for a DenseNet-169 teacher. Similar to the results in Tab. 2, we find that e^{2}KD adds significant gains to 'vanilla' KD across dataset sizes (50 Shots, 200 Shots, full data) and, as it does not rely on matching specific blocks between architectures (cf. [8,39]), it seamlessly works across architectures. Further, e^{2}KD can also be used with 'frozen' (❄) explanations by augmenting images and pre-computed explanations jointly (Sec. 3.3).

B-cos Models	**50 Shots**		**200 Shots**		**Full data**	
Teacher DenseNet-169 Accuracy 75.2%	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.
Baseline ResNet-18	32.6	34.5	53.9	58.4	68.7	75.5
KD [4,17]	37.3	40.2	51.3	55.6	71.2	78.8
+ **e^{2}KD (B-cos)**	**45.4**	**49.0**	**55.7**	**60.7**	**71.9**	**79.8**
	+8.1	**+8.8**	**+4.4**	**+5.1**	**+0.7**	**+1.0**
❄ KD	33.4	35.7	50.4	54.5	68.7	75.2
❄ + **e^{2}KD (B-cos)**	**38.7**	**41.7**	**53.6**	**58.3**	**69.5**	**76.4**
	+5.3	**+6.0**	**+3.2**	**+3.8**	**+0.8**	**+1.2**

4.2 e^{2}KD Improves Learning the 'Right' Features

Setup. To assess whether the students learn to use the same input features as the teacher (Sec. 3.2 Desideratum 2), we use the Waterbirds-100 dataset [28], a binary classification task between land- and waterbirds, in which birds are highly correlated with the image backgrounds during training. As teachers, we use pre-trained ResNet-50 models from [24], which were guided to use the bird features instead of the background; as in Sec. 4.1, we use conventional and B-cos models and provide results obtained via prior work for reference.We further demonstrate the model-agnostic aspect of e^{2}KD by testing a variety of CNN architectures as students. In light of the findings by [4] that long teaching schedules and strong data augmentations help, we explore three settings[1]: (1) 700 epochs, (2) with add. mixup [41], as well as (3) training 5x longer ('patient teaching').

Results. In Fig. 2, we present our results on the Waterbirds for standard models (see appx. B.2 for B-cos models). We evaluate the accuracy and student-teacher agreement of each method on object-background combinations not seen during training (i.e. 'Waterbird on Land' & 'Landbird on Water') to see how well the students learned from the teacher to rely on the 'right' input features (i.e. birds).

Across all settings, e^{2}KD significantly boosts the out-of-distribution performance of KD on both accuracy and agreement. Despite its simplicity, it compares favourably to prior work, indicating that e^{2}KD indeed promotes faithful distillation. Notably, Fig. 2 is also an example of how the in-distribution performance of KD methods may not fully reflect their differences.We also find clear qualitative improvements in the explanations focusing on the 'right' features, see Fig. 3 for B-cos models and Sec. A.1 in the appendix for standard models.

[1] Compared to ImageNet, the small size of the Waterbirds-100 dataset allows for reproducing the 'patient teaching' results with limited compute.

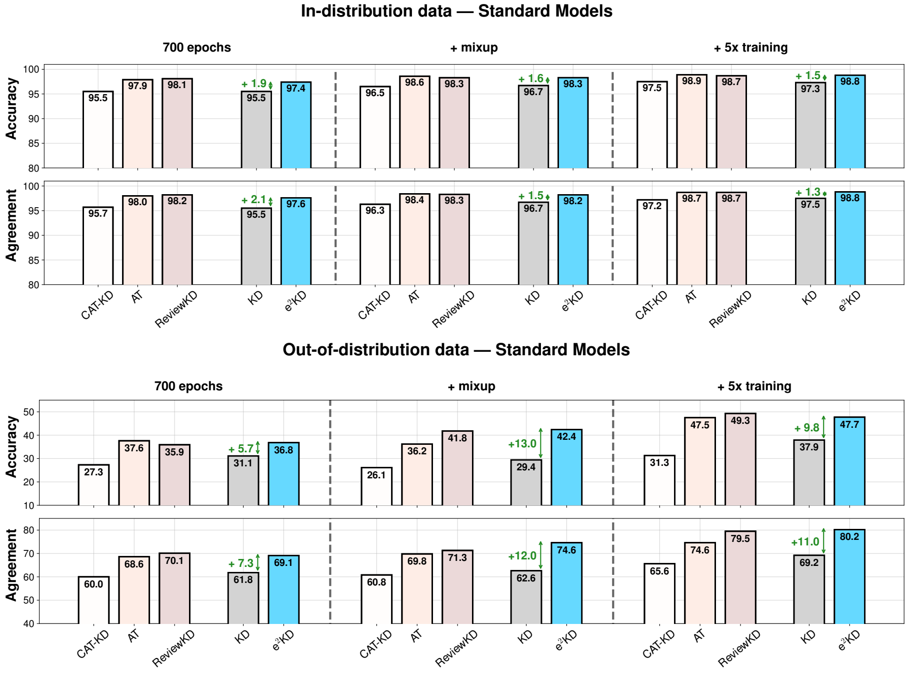


Fig. 2: KD for standard models on Waterbirds-100. We show the accuracy and agreement on in-distribution (**top**) and out-of-distribution (**bottom**) test samples when distilling from a ResNet-50 teacher to a ResNet-18 student with various KD approaches. Following [4], we additionally evaluate the effectiveness of adding mixup (**col. 2**) and, additionally, long teaching (**col. 3**). We find that our proposed e^{2}KD provides significant benefits over vanilla KD, and is further enhanced under long teaching and mixup. We show the performance of prior work for reference, and find that e^{2}KD performs competitively. For results on B-cos models, see appendix B.2 and Fig. 3.

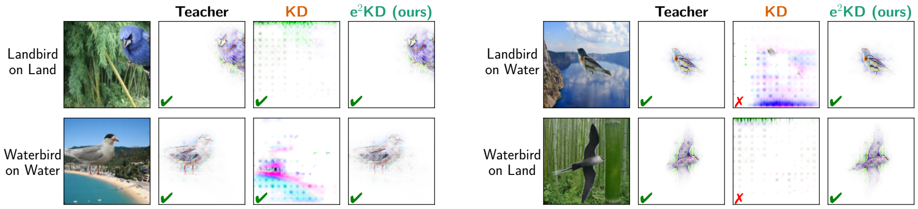


Fig. 3: Comparing explanations for KD on Waterbirds. Here we visualize B-cos explanations, when distilling a B-cos ResNet-50 teacher (**col. 2**) to a B-cos ResNet-18 student with KD (**col. 3**) and e^{2}KD (**col. 4**). While for in-distribution data (**left**) the different focus of the models (foreground/background) does not affect the models' predictions (correct predictions marked by ✓), it results in wrong predictions under distribution shift (**right**, incorrect predictions marked by ✗). For additional qualitative results, including standard models with GradCAM explanations, see appendix A.1.

Further, consistent with [4], we find mixup augmentation and longer training schedules to also significantly improve agreement. This provides additional evidence for the hypothesis put forward by [4] that KD *could* be sufficient for function matching if performed for long enough. As such, and given the simplicity of the dataset, the low resource requirements, and a clear target (100% agreement on unseen combinations), we believe the waterbirds dataset to constitute a great benchmark for future research towards faithful KD.

Lastly, given that e^{2}KD does not make any reference to model architecture and simply matches the explanations on top of KD, we find that it consistently improves out-of-distribution performance across different student architectures, see Tab. 4. As we discuss in the next section, the model-agnostic nature of e^{2}KD also seamlessly allows to transfer knowledge between CNNs and ViTs.

Table 4: Out-of-distribution results on Waterbirds-100 across student architectures. We show accuracy and agreement results on out-of-distribution samples when distilling a standard ResNet-50 teacher (similar to Fig. 2) to different students. e^{2}KD results in consistent gains across students, by simply matching the explanations.

Method	ConvNext		EfficientNet		MobileNet		ShuffleNet	
	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.
KD	20.5	55.5	27.5	59.0	22.3	57.0	23.1	57.1
+ e^{2}KD (GradCAM)	**32.2**	**64.4**	**37.8**	**68.7**	**36.0**	**68.2**	**37.0**	**68.6**

4.3 e^{2}KD Improves the Student's Interpretability

In this section, we present results on maintaining the teacher's interpretability (cf. Sec. 3.2 Desideratum 3). In particular, we show that e^{2}KD naturally lends itself to distilling localization properties of the teacher into the students (Sec. 4.3 'Distill on VOC') and that even architectural priors of a CNNs can be transferred to ViT students (Sec. 4.3 'Distill to ViT').

Distill on VOC. We assess how well the focused explanations are preserved. **Setup.** To assess whether the students learn to give similar explanations as the teachers, we distill B-cos ResNet-50 teachers into B-cos ResNet-18 students on PASCAL VOC [11] in a multi-label classification setting. Specifically, we use two different teachers from [24]: one with explanations of high EPG [34] score (EPG Teacher), and one with explanations of high IoU score (IoU Teacher). To quantify the students' focus, we then measure the EPG and IoU scores of the explanations with respect to the dataset's bounding box annotations in a multi-label classification setting. As these teachers are trained explicitly to exhibit certain properties in their explanations, a *faithfully distilled* student should optimally exhibit the same properties.
Results. As we show in Tab. 5, the explanations of an e^{2}KD student indeed more closely mirror those of the teacher than a student trained via vanilla KD:

Table 5: (Left) e^{2}KD results on VOC. We compare KD and e^{2}KD when distilling from a B-cos ResNet-50 teacher guided [24] to either optimize for EPG (*left*) or IoU (*right*). Explanations of the e^{2}KD student better align with those of the teacher, as evidenced by significantly higher EPG (IoU) scores when distilled from the EPG (IoU) teacher. e^{2}KD students also achieve higher accuracy (F1).**(Right) KD on unrelated images**. A B-cos DenseNet-169 teacher model, *left:* trained on the SUN [35] is distilled with ImageNet (IMN→SUN), and *right:* trained on ImageNet is distilled with SUN (SUN→IMN). In both cases, the B-cos ResNet-18 student distilled with e^{2}KD achieves significantly higher accuracy and agreement scores than student trained via vanilla KD.

	KD on the VOC Dataset						**KD on Unrelated Images**			
	EPG Teacher			**IoU Teacher**			**IMN→SUN**		**SUN→IMN**	
	EPG	IoU	F1	EPG	**IoU**	F1	Acc.	Agr.	Acc.	Agr.
Teacher	75.7	21.3	72.5	65.0	49.7	72.8	60.5	-	75.2	-
Baseline	50.0	29.0	58.0	50.0	29.0	58.0	57.7	67.9	68.7	75.5
KD	60.1	31.6	60.1	58.9	35.7	62.7	53.5	65.0	14.9	16.7
+ **e^{2}KD**	**71.1**	24.8	**67.6**	60.3	**45.7**	**64.8**	**54.9**	**67.7**	**19.8**	**22.1**

e^{2}KD students exhibit significantly higher EPG when distilled from the EPG teacher (EPG: 71.1 vs. 60.3) and vice versa (IoU: 45.7 vs. 24.8). In contrast, 'vanilla' KD students show only minor differences (EPG: 60.1 vs. 58.9; IoU: 35.7 vs. 31.6). These improvements also show qualitatively (Fig. 4), with the e^{2}KD students reflecting the teacher's focus much more faithfully in their explanations.

While this might be expected as e^{2}KD explicitly optimizes for explanation similarity, we would like to highlight that this not only ensures that the desired properties of the teachers are better represented in the student model, but also significantly improves the students' performance (e.g., F1: 60.1→67.6 for the EPG teacher). As such, we find e^{2}KD to be an easy-to-use and effective addition to vanilla KD for improving both interpretability as well as task performance.

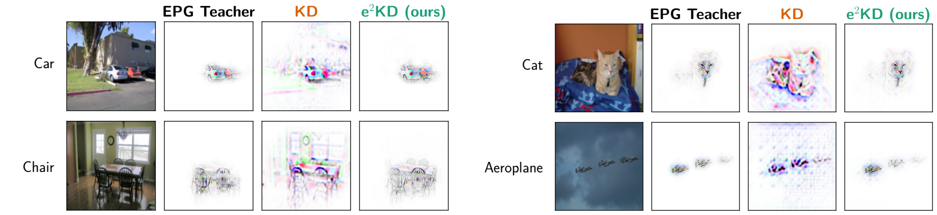
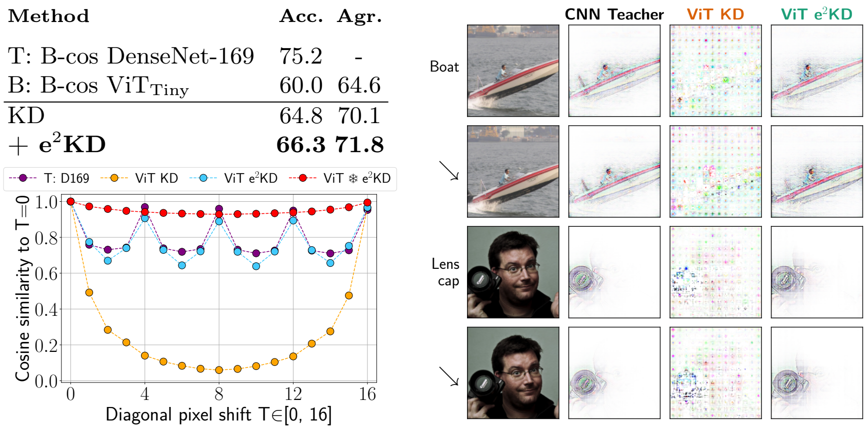


Fig. 4: Maintaining focused explanations. We visualize B-cos explanations, when distilling a B-cos ResNet-50 teacher that has been trained to not focus on confounding input features (**col. 2**), to a B-cos ResNet-18 student with KD (**col. 3**) and e^{2}KD (**col. 4**). Explanations of e^{2}KD students are significantly closer to the teacher's (and hence more human-aligned). Samples are drawn from the VOC test set, with all models correctly classifying the shown samples. For more qualitative results, see appendix A.2.



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
+ e^{2}KD	**66.3**	**71.8**

Fig. 5: Distilling inductive biases (CNN→ViT). We distill a B-cos DenseNet-169 teacher to a B-cos ViT$_{\text{Tiny}}$. **Top-Left:** e^{2}KD yields significant gains in accuracy and agreement. **Bottom-Left:** Cosine similarity of explanations for shifted images w.r.t. the unshifted image (T=0). With e^{2}KD (blue) the ViT student learns to mimic the shift periodicity of the teacher (purple), despite the inherent periodicity of 16 of the ViT architecture (seen for vanilla KD, yellow). Notably, e^{2}KD with frozen explanations yields shift-equivariant students (red), see also Sec. 4.3 'Distill to ViT'. **Right:** e^{2}KD significantly improves the explanations of the ViT model, thus maintaining the utility of the explanations of the CNN teacher model. While the explanations for KD change significantly under shift (subcol. 3), for e^{2}KD (subcol. 4), as with the CNN teacher (subcol. 2), the explanations remain consistent. See also appendix A.3.

Distill to ViT. We assess if inductive biases of CNN can be distilled to ViT. **Setup.** To test whether students learn architectural priors of the models, we evaluate whether a B-cos ViT$_{\text{Tiny}}$ student can learn to give explanations that are similar to those of a pretrained CNN (B-cos DenseNet-169) teacher model; for this, we again use the ImageNet dataset.

Results. In line with the results of the preceding sections, we find (Fig. 5, left) that e^{2}KD significantly improves the accuracy of the ViT student model (64.8→66.3), as well as the agreement with the teacher (70.1→71.8).

Interestingly, we find that the ViT student's explanations seem to become similarly robust to image shifts as those of the teacher (Fig. 5, bottom-left and right.). Specifically, note that the image tokenization of the ViT model using vanilla KD (extracting non-overlapping patches of size 16×16) induces a periodicity of 16 with respect to image shifts T, see, e.g., Fig. 5 (bottom-left, yellow curve): here, we plot the cosine similarity of the explanations[2] at various shifts with respect to the explanation given for the original, unshifted image (T=0). In contrast, due to smaller strides (stride$\in\{1,2\}$ for any layer) and overlapping convolutional kernels, the CNN teacher model is inherently more robust to image shifts, see Fig. 5 (purple curve), exhibiting a periodicity of 4. A ViT student

[2] We compute the similarity of the intersecting area of the explanations.

trained via e^{2}KD learns to mimic the behaviour of the teacher (see Fig. 5, blue curve) and exhibits the same periodicity, indicating that e^{2}KD indeed helps the student learn a function more similar to the teacher.

In Fig. 5 (right), we see that e^{2}KD also significantly improves the explanations of the ViT model. We show explanations for original and 8-pixel diagonally shifted (↘) images. Our ViT's explanations are more robust to shifts and more interpretable, thus maintaining the utility of the explanations of the teacher.

4.4 e^{2}KD with Frozen Explanations

In the previous sections, we showed that e^{2}KD is a robust approach that provides consistent gains even when only limited data is available (see Sec. 4.1) and works across different architectures (e.g., DenseNet→ResNet or DenseNet→ViT, see Secs. 4.1 and 4.3 'Distill to ViT'). In the following, we show that e^{2}KD even works when only 'approximate' explanations for the teacher are available (cf. Sec. 3.3).
Setup. To test the robustness of e^{2}KD when using frozen explanations, we distill from a B-cos DenseNet-169 teacher to a B-cos ResNet-18 student using pre-computed, frozen explanations on the ImageNet dataset. We also evaluate across varying dataset sizes, as in Sec. 4.1.
Results. Tab. 3 (bottom) shows that e^{2}KD with frozen explanations is effective for improving both the accuracy and agreement over KD with frozen logits across dataset sizes (e.g. accuracy: 33.4→38.7 for 50 shots). Furthermore, e^{2}KD with frozen explanations also outperforms vanilla KD under both metrics when using limited data (e.g. accuracy: 37.3→38.7 for 50 shots). As such, a frozen teacher constitutes a more cost-effective alternative for obtaining the benefits of e^{2}KD, whilst also highlighting its robustness to using 'approximate' explanations.

Our results also indicate that it might be possible to instill desired properties into a DNN model even beyond knowledge distillation. Note that the frozen explanations are *by design* equivariant explanations across shifts and crops. Based on our observations for the ViTs (cf. Sec. 4.3), we thus expect a student trained on frozen explanations to become almost *fully shift-equivariant*, which is indeed the case for our ViT students (see Fig. 5, bottom-left, red curve, ViT ❄ e^{2}KD).

5 Conclusion

We proposed a simple approach to promote the faithfulness of knowledge distillation (KD) by explicitly optimizing for the explanation similarity between the teacher and the student, and showed its effectiveness in distilling the teacher's properties under multiple settings. Specifically, e^{2}KD helps the student (1) achieve competitive and often higher accuracy and agreement than vanilla KD, (2) learn to be 'right for the right reasons', and (3) learn to give similar explanations as the teacher, e.g. even when distilling from a CNN teacher to a ViT student. Finally, we showed that e^{2}KD is robust in the presence of limited data, approximate explanations, and across model architectures. In short, we find e^{2}KD to be a simple but versatile addition to KD that allows for a more faithful distillation of the teacher, whilst also maintaining competitive task performance.

References

1. Alharbi, R., Vu, M.N., Thai, M.T.: Learning Interpretation with Explainable Knowledge Distillation. In: 2021 IEEE International Conference on Big Data (Big Data). pp. 705–714. IEEE Computer Society, Los Alamitos, CA, USA (dec 2021)
2. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On Pixel-wise Explanations for Non-linear Classifier Decisions by Layer-wise Relevance Propagation. PloS one **10**(7), e0130140 (2015)
3. Bansal, G., Nushi, B., Kamar, E., Weld, D.S., Lasecki, W.S., Horvitz, E.: Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. In: AAAI. vol. 33, pp. 2429–2437 (2019)
4. Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R., Kolesnikov, A.: Knowledge Distillation: A Good Teacher Is Patient and Consistent. In: CVPR (June 2022)
5. Böhle, M., Fritz, M., Schiele, B.: B-cos Networks: Alignment Is All We Need for Interpretability. In: CVPR (June 2022)
6. Böhle, M., Singh, N., Fritz, M., Schiele, B.: B-cos Alignment for Inherently Interpretable CNNs and Vision Transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers. In: ICCV. pp. 9650–9660 (2021)
8. Chen, P., Liu, S., Zhao, H., Jia, J.: Distilling Knowledge via Knowledge Review. In: CVPR (June 2021)
9. Council of the European Union: Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206` (2021), accessed: 2023-11-15
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR. IEEE (2009)
11. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. IJCV **88** (2009)
12. Faghri, F., Pouransari, H., Mehta, S., Farajtabar, M., Farhadi, A., Rastegari, M., Tuzel, O.: Reinforce Data, Multiply Impact: Improved Model Accuracy and Robustness with Dataset Reinforcement. In: ICCV (October 2023)
13. Gao, Y., Sun, T.S., Bai, G., Gu, S., Hong, S.R., Liang, Z.: RES: A Robust Framework for Guiding Visual Explanation. In: KDD. pp. 432–442 (2022)
14. Gao, Y., Sun, T.S., Zhao, L., Hong, S.R.: Aligning Eyes between Humans and Deep Neural Network through Interactive Attention Alignment. Proceedings of the ACM on Human-Computer Interaction **6**(CSCW2), 1–28 (2022)
15. Guo, Z., Yan, H., Li, H., Lin, X.: Class Attention Transfer Based Knowledge Distillation. In: CVPR (2023)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR. pp. 770–778 (2016)
17. Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network. In: NeurIPSW (2015)
18. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely Connected Convolutional Networks. In: CVPR. pp. 4700–4708 (2017)
19. Liu, X., Li, L., Li, C., Yao, A.: Norm: Knowledge distillation via n-to-one representation matching (2023), `https://arxiv.org/abs/2305.13803`

20. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: ICLR (2019)
21. Ojha, U., Li, Y., Lee, Y.J.: What Knowledge gets Distilled in Knowledge Distillation? arXiv preprint arXiv:2205.16004 (2022)
22. Petryk, S., Dunlap, L., Nasseri, K., Gonzalez, J., Darrell, T., Rohrbach, A.: On Guiding Visual Attention with Language Specification. CVPR (2022)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models from Natural Language Supervision. In: ICML. pp. 8748–8763 (2021)
24. Rao, S., Böhle, M., Parchami-Araghi, A., Schiele, B.: Studying How to Efficiently and Effectively Guide Models with Explanations. In: ICCV (2023)
25. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In: KDD. pp. 1135–1144 (2016)
26. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: FitNets: Hints for Thin Deep Nets. In: Bengio, Y., LeCun, Y. (eds.) ICLR (2015)
27. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. In: IJCAI (2017)
28. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In: ICLR (2020)
29. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: ICCV (2017)
30. Shen, Z., Xing, E.: A Fast Knowledge Distillation Framework for Visual Recognition. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV. pp. 673–690. Springer Nature Switzerland, Cham (2022)
31. Srinivas, S., Fleuret, F.: Knowledge Transfer with Jacobian Matching. In: ICML (2018)
32. Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A.A., Wilson, A.G.: Does Knowledge Distillation Really Work? In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) NeurIPS. vol. 34. Curran Associates, Inc. (2021)
33. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: International Conference on Learning Representations (2020)
34. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-cam: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2020)
35. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN database: Large-Scale Scene Recognition from Abbey to Zoo. In: CVPR. pp. 3485–3492 (June 2010)
36. Yan, S., Xiong, Y., Kundu, K., Yang, S., Deng, S., Wang, M., Xia, W., Soatto, S.: Positive-congruent Training: Towards Regression-free Model Updates. In: CVPR. pp. 14299–14308 (2021)
37. Yang, P., Xie, M.K., Zong, C.C., Feng, L., Niu, G., Sugiyama, M., Huang, S.J.: Multi-Label Knowledge Distillation. In: ICCV (2023)
38. Yun, S., Oh, S.J., Heo, B., Han, D., Choe, J., Chun, S.: Re-Labeling ImageNet: From Single to Multi-Labels, From Global to Localized Labels. In: CVPR. pp. 2340–2350 (June 2021)
39. Zagoruyko, S., Komodakis, N.: Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In: ICLR (2017)

40. Zeyu, D., Yaakob, R., Azman, A., Mohd Rum, S.N., Zakaria, N., Ahmad Nazri, A.S.: A grad-cam-based knowledge distillation method for the detection of tuberculosis. In: 2023 International Conference on Information Management (ICIM). pp. 72–77 (2023)
41. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond Empirical Risk Minimization. In: ICLR (2018)
42. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: CVPR (2022)
43. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning Deep Features for Discriminative Localization. In: CVPR (2016)