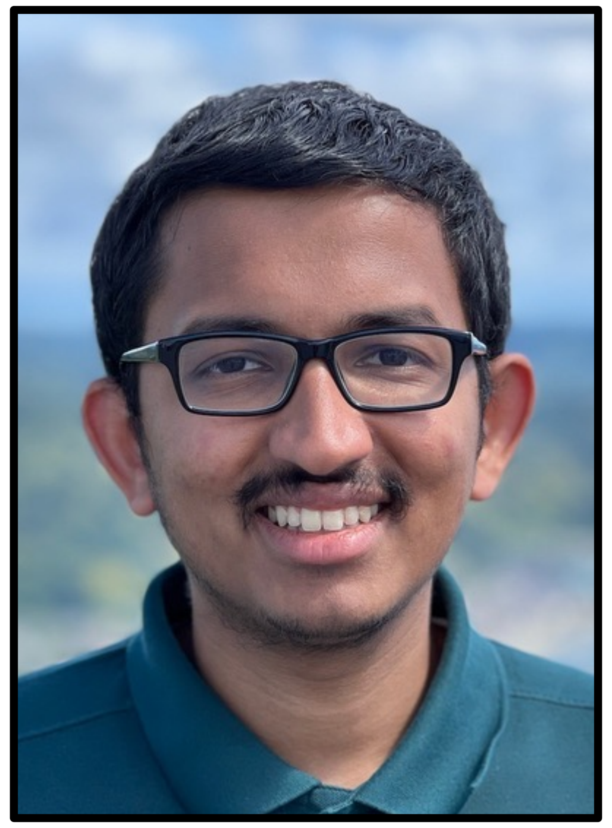
Good Teachers Explain: Explanation-enhanced Knowledge Distillation


Amin Parchami-Araghi


Moritz Böhle


Sukrut Rao


Bernt Schiele

Max Planck Institute for Informatics, Saarland Informatics Campus

Knowledge Distillation

Simply match the logits between teacher and student for every input.

$$D_{\mathrm{KL}}(P^T \| P^S) = \sum_{j=1}^{C} P_j^T(x) \log \left(\frac{P_j^T(x)}{P_j^S(x)} \right)$$

Goal: a student with the same accuracy as the teacher

Recently: For long-enough distillation, the student can reach teacher's accuracy.

But does this indicate a successful distillation?

[1] Beyer et al., Knowledge distillation: A good teacher is patient and consistent, CVPR 2022

Knowledge Distillation

Besides accuracy, a recent work [2] evaluate the `agreement' between the two models.

Test Samples

Teacher: ✅ ✅ ✅ ✅ ✅ ❌ ❌ ✅ ❌

Student: ❌ ✅ ❌ ❌ ✅ ✅ ✅ ✅ ✅

Both 66% Accurate, only 33% agreement.

Despite matching accuracies, the agreement can be significantly lower.

*There can be a **disparity between the two functions** despite having same accuracy*

[2] Stanton et al., Does Knowledge Distillation Really Work?, NeurIPS 2021

Our Work: Faithful Knowledge Distillation

We extend the work of Stanton et al. and aim towards ***faithful KD.***

Faithful KD looks beyond accuracy, aiming for *functionally similar* Teacher and Student

[2] Stanton et al., Does Knowledge Distillation Really Work?, NeurIPS 2021

Our Work: Faithful Knowledge Distillation

We extend the work of Stanton et al. and aim towards ***faithful KD.***

Faithful KD looks beyond accuracy, aiming for *functionally similar* Teacher and Student

This implies:

- High teacher-student agreement

  Especially under limited-data settings.

Test Samples

Teacher: ✅ ✅ ✅ ✅ ✅ ❌ ❌ ✅ ❌

Student: ❌ ✅ ❌ ❌ ✅ ✅ ✅ ✅ ✅

[2] Stanton et al., Does Knowledge Distillation Really Work?, NeurIPS 2021

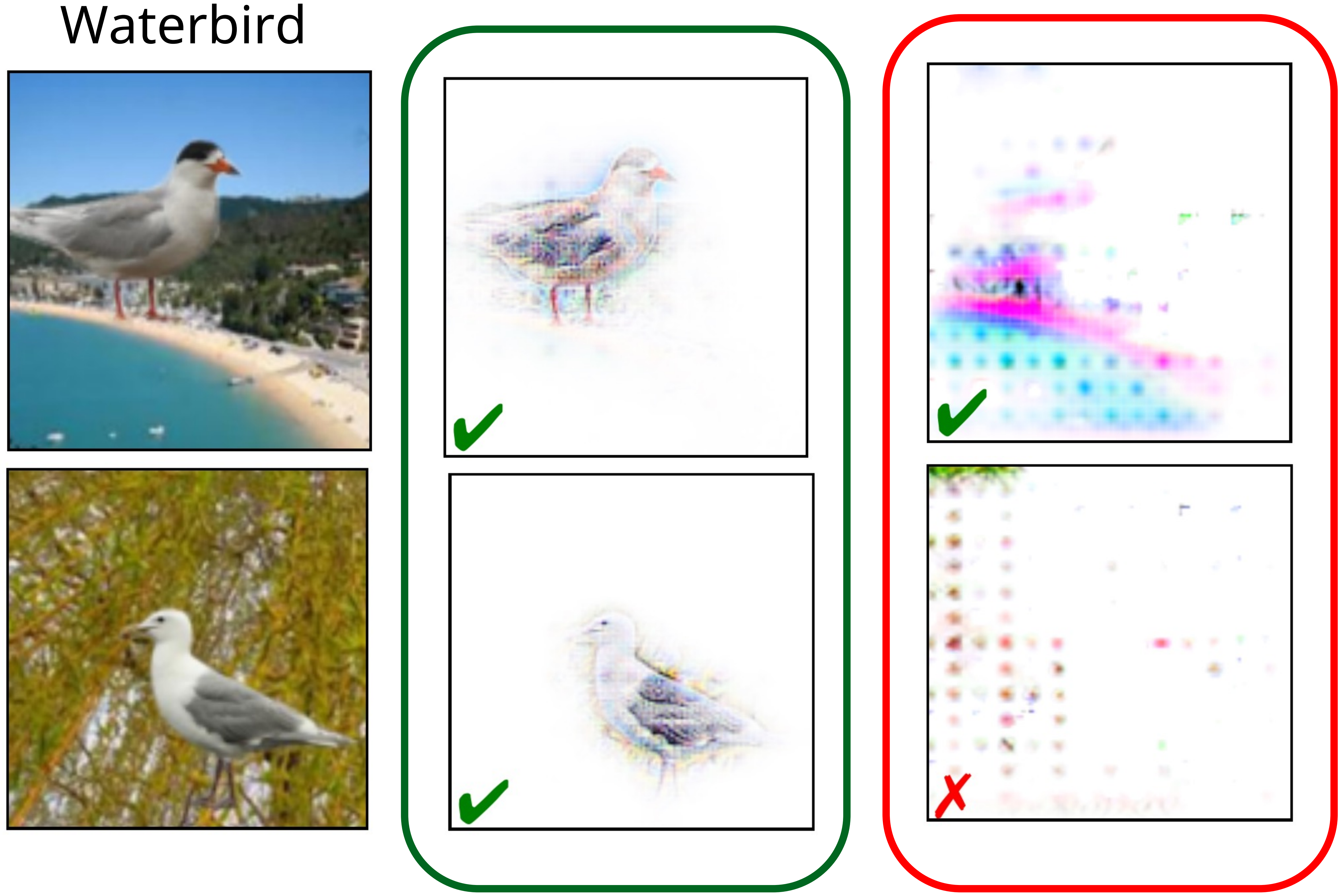
Our Work: Faithful Knowledge Distillation

We extend the work of Stanton et al. and aim towards ***faithful KD.***

Faithful KD looks beyond accuracy, aiming for *functionally similar* Teacher and Student

This implies:

- High teacher-student agreement
- Similar predictions, *but for similar reasons*



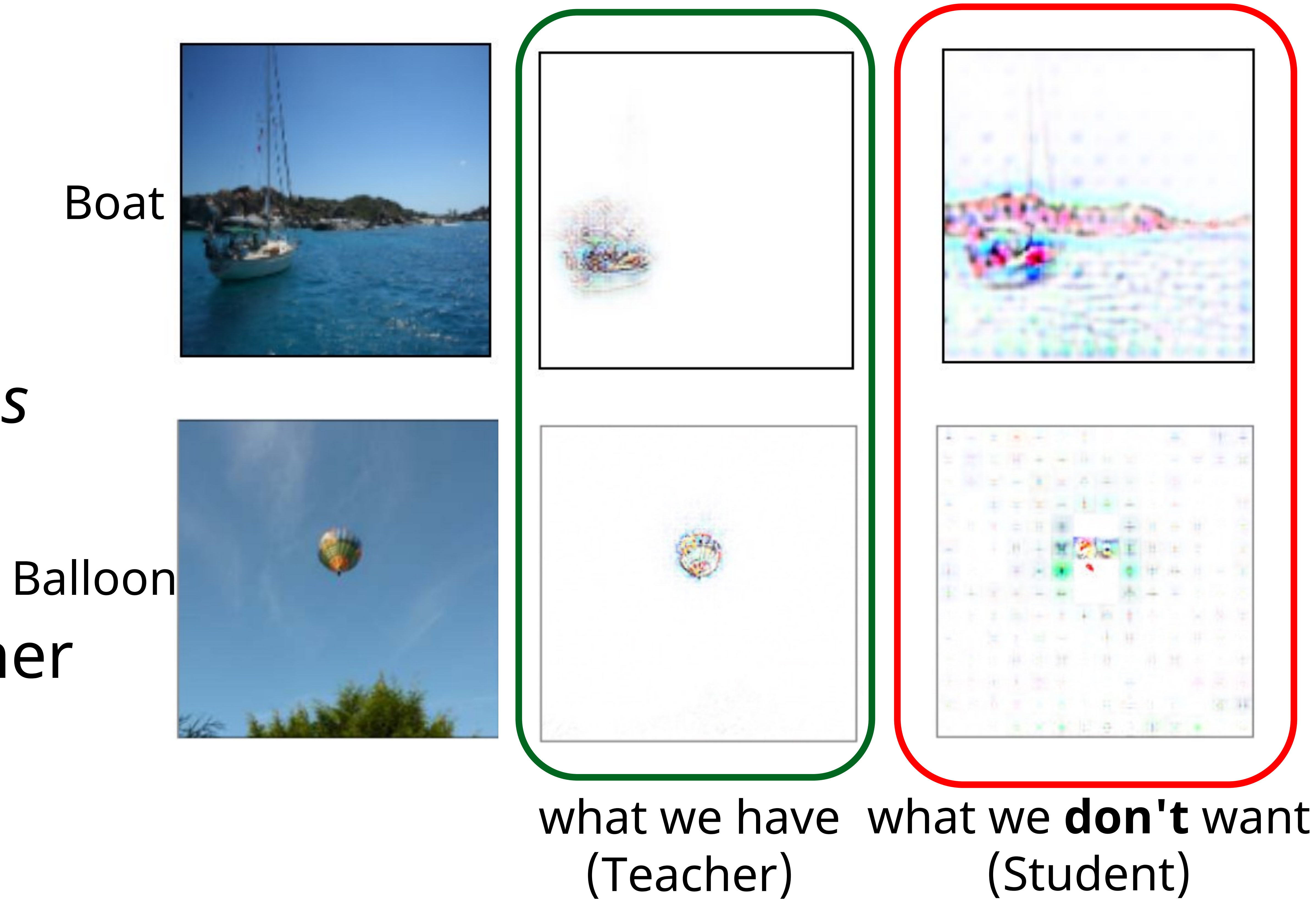
Our Work: Faithful Knowledge Distillation

We extend the work of Stanton et al. and aim towards ***faithful KD.***

Faithful KD looks beyond accuracy, aiming for *functionally similar* Teacher and Student

This implies:

- High teacher-student agreement
- Similar predictions, *but for similar reasons*
- Maintain the interpretability of the teacher



Our Work: Faithful Knowledge Distillation

We extend the work of Stanton et al. and aim towards ***faithful KD.***

Faithful KD looks beyond accuracy, aiming for *functionally similar* Teacher and Student

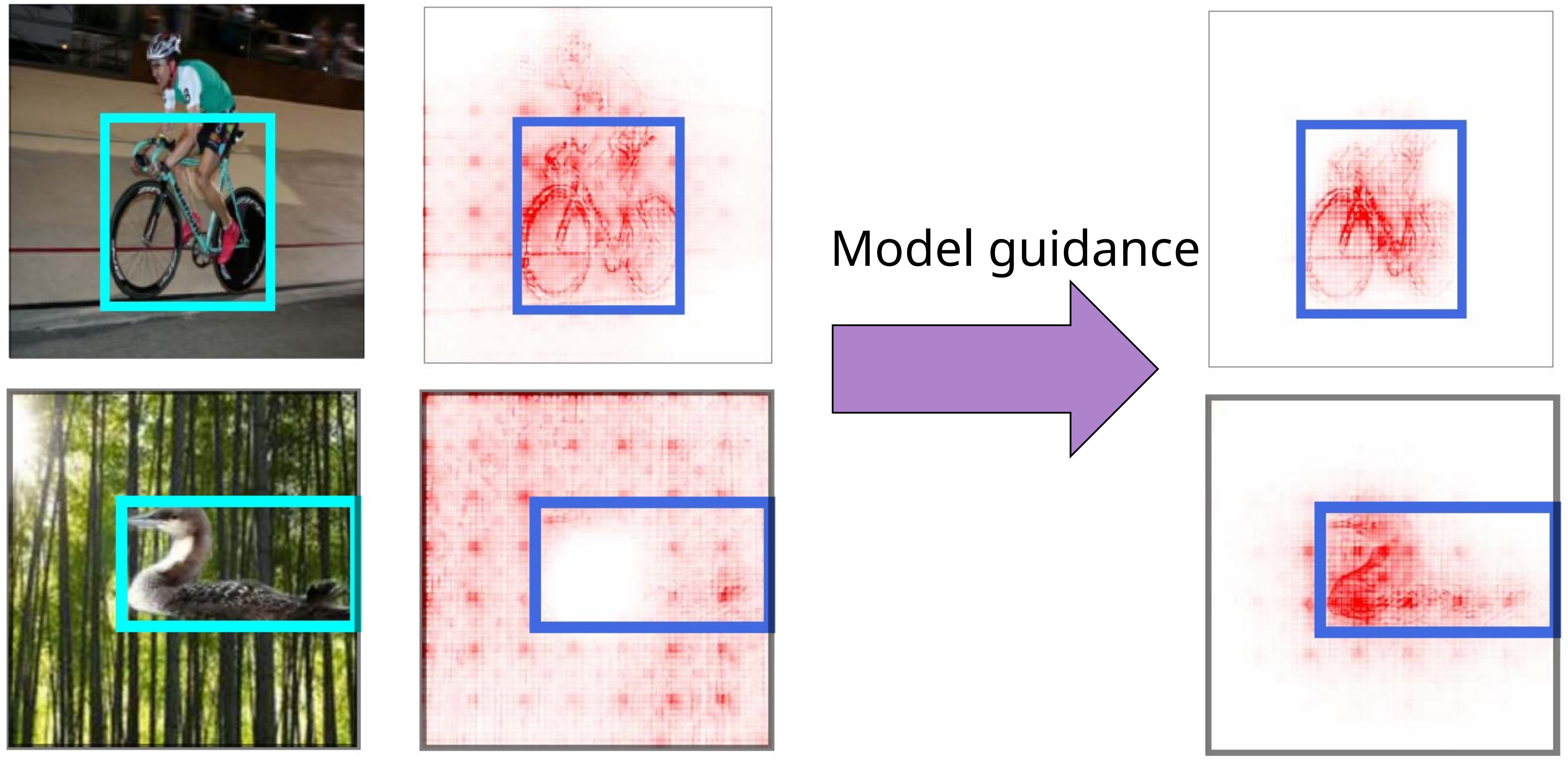
This implies:

- High teacher-student agreement
- Similar predictions, *but for similar reasons*
- Maintain the interpretability of the teacher

three desiderata and settings for each

Our Work: Leverage explanation methods!

- We want to make the two models more functionally similar!
- Existing explanation methods have shown to be powerful for steering models [3]



Can we simply use existing explanation methods for a more faithful KD?

[3] Rao et. al. Studying How to Efficiently and Effectively Guide Models using Explanations, ICCV 2023

Our Work: Optimizing for Faithfulness

Inspired by model guidance,

we explore the benefits of simply optimizing for explanation similarity

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

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

Our Work: Optimizing for Faithfulness

Inspired by model guidance,

we explore the benefits of simply optimizing for explanation similarity

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

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

- Label- and Parameter-free
- Model-agnostic
- Utilize existing explanation methods

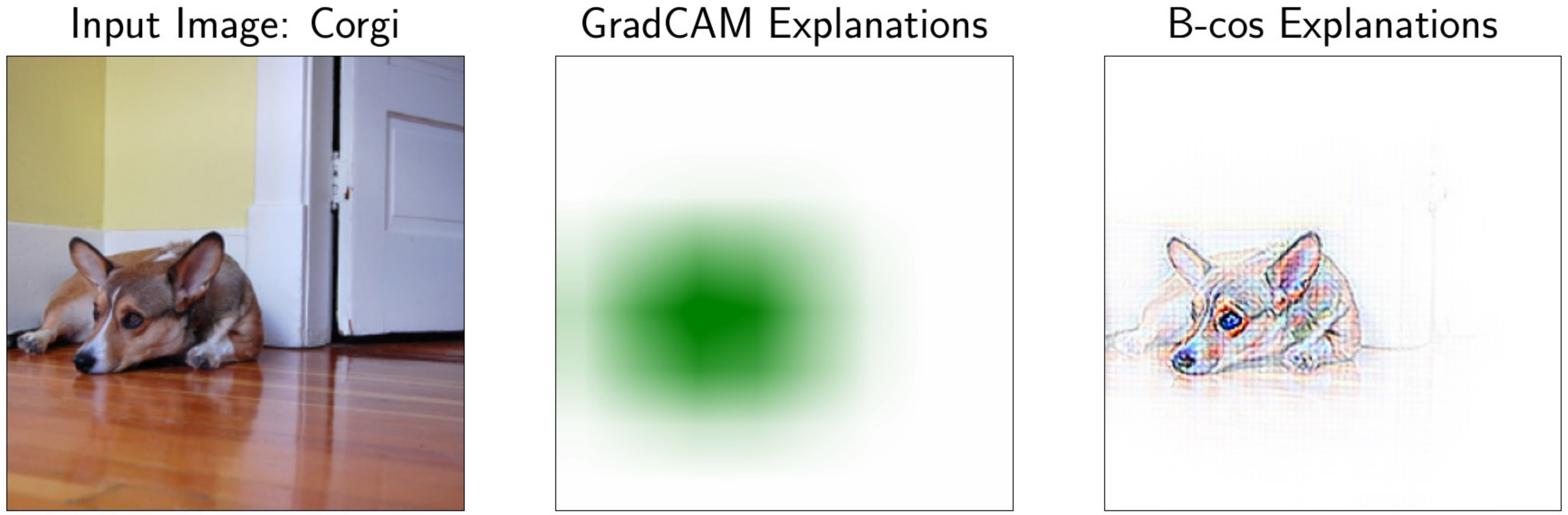
Our Work: Optimizing for Faithfulness

Inspired by model guidance,
we explore the benefits of simply optimizing for explanation similarity

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

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

- Label- and Parameter-free
- Model-agnostic
- Utilize existing explanation methods



Selvaraju et al., Grad-CAM, ICCV 2017; Böhle et al., B-cos, CVPR 2022 & TPAMI 2024

Desideratum 1: High Agreement with Teacher

Setting: ImageNet;
Distill on different amounts of available data

Evaluating on the complete test set

Larger gains for smaller distillation sizes

Standard Models Teacher ResNet-34 Accuracy 73.3%	**50 Shots**		**200 Shots**		**Full data**	
	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.
KD [37, 5]	49.8	55.5	63.1	71.9	**71.8**	81.2
+ e²KD (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

B-cos Models Teacher ResNet-34 Accuracy 72.3%	**50 Shots**		**200 Shots**		**Full data**	
	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.
KD [37, 5]	35.3	38.4	56.5	62.9	70.3	79.9
+ e²KD (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

B-cos Models Teacher DenseNet-169 Accuracy 75.2%	**50 Shots**		**200 Shots**		**Full data**	
	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.
KD [37, 5]	37.3	40.2	51.3	55.6	71.2	78.8
+ e²KD (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

Desideratum 1: High Agreement with Teacher

Distillation: B-cos DenseNet-169 → B-cos ResNet-18

	SUN397 Teacher / SUN397 Student / ImageNet images		ImageNet Teacher / ImageNet Student / SUN397 images	
Distill: / To: / With:	Acc.	Agr.	Acc.	Agr.
Teacher DenseNet-169	60.5	-	75.2	-
Baseline ResNet-18	57.7	67.9	68.7	75.5
KD [4, 19]	53.5	65.0	14.9	16.7
+ e^2KD (B-cos)	**54.9**	**67.7**	**19.8**	**22.1**

e^2KD provides gains even on unrelated images

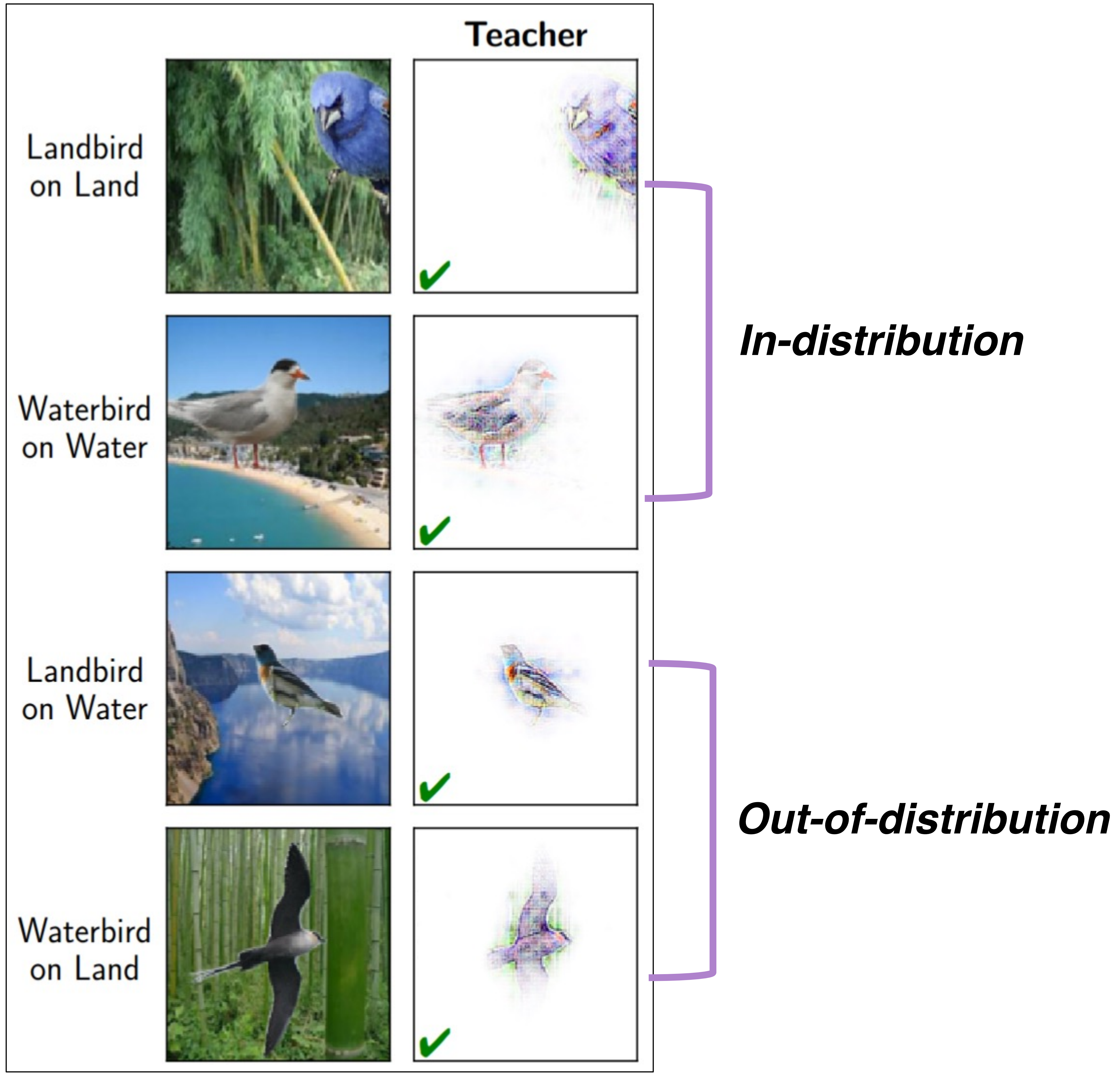
Desideratum 2: Learning the 'Right' Features

Task: Classify Landbird vs. Waterbird

Distillation Data: Landbird on Land, Waterbird on Water

Test-time: Landbird on Water, Waterbird on Land

Teacher: ResNet-50 explicitly guided to focus on the bird



Focusing on the 'Right' input features gives OOD robustness.

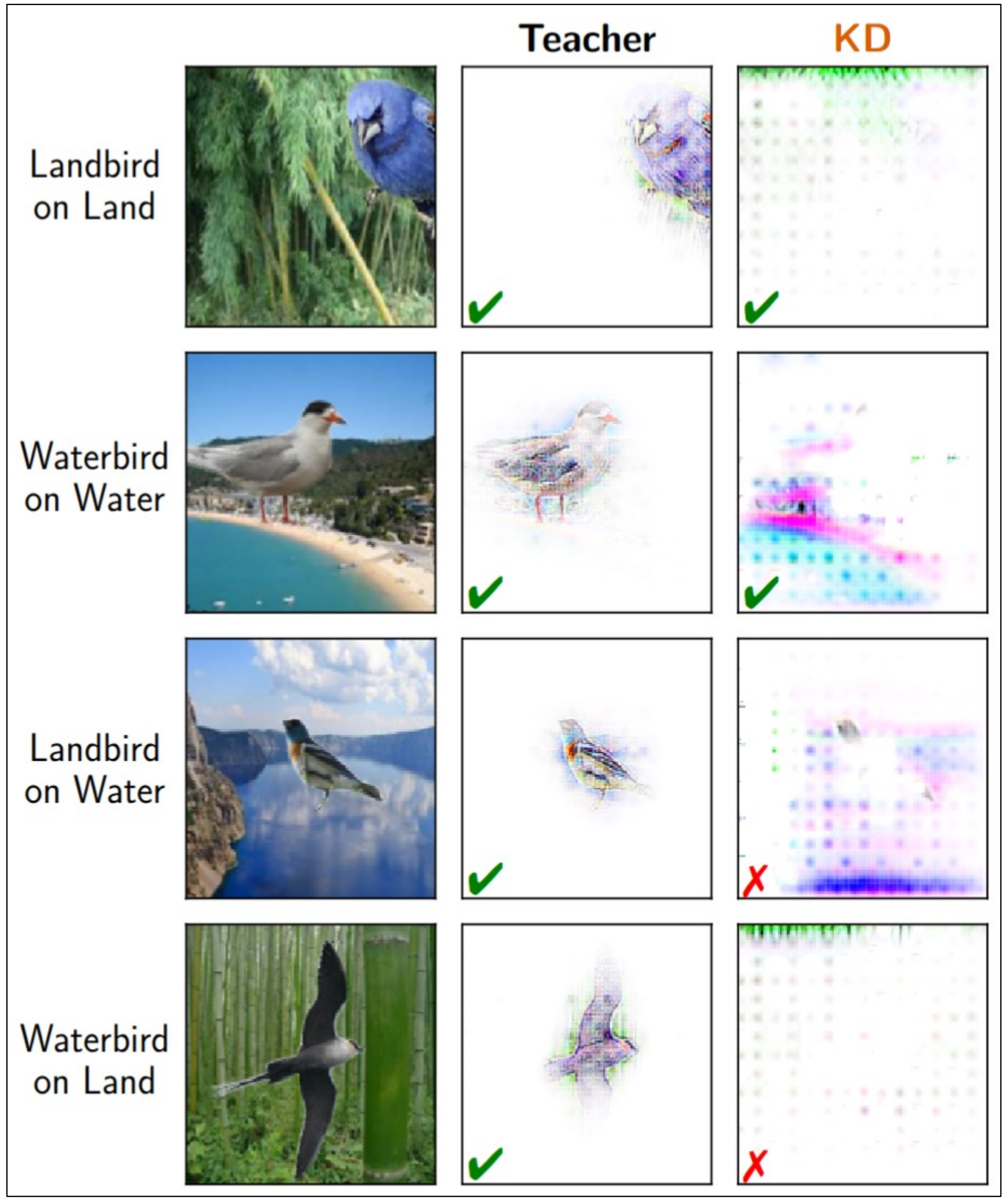
Desideratum 2: Learning the 'Right' Features

Task: Classify Landbird vs. Waterbird

Distillation Data: Landbird on **Land**, Waterbird on **Water**

Test-time: Landbird on **Water**, Waterbird on **Land**

Teacher: ResNet-50 explicitly guided to focus on the bird



Focusing on the 'Right' input features gives OOD robustness.

The student might deviate from the teacher!

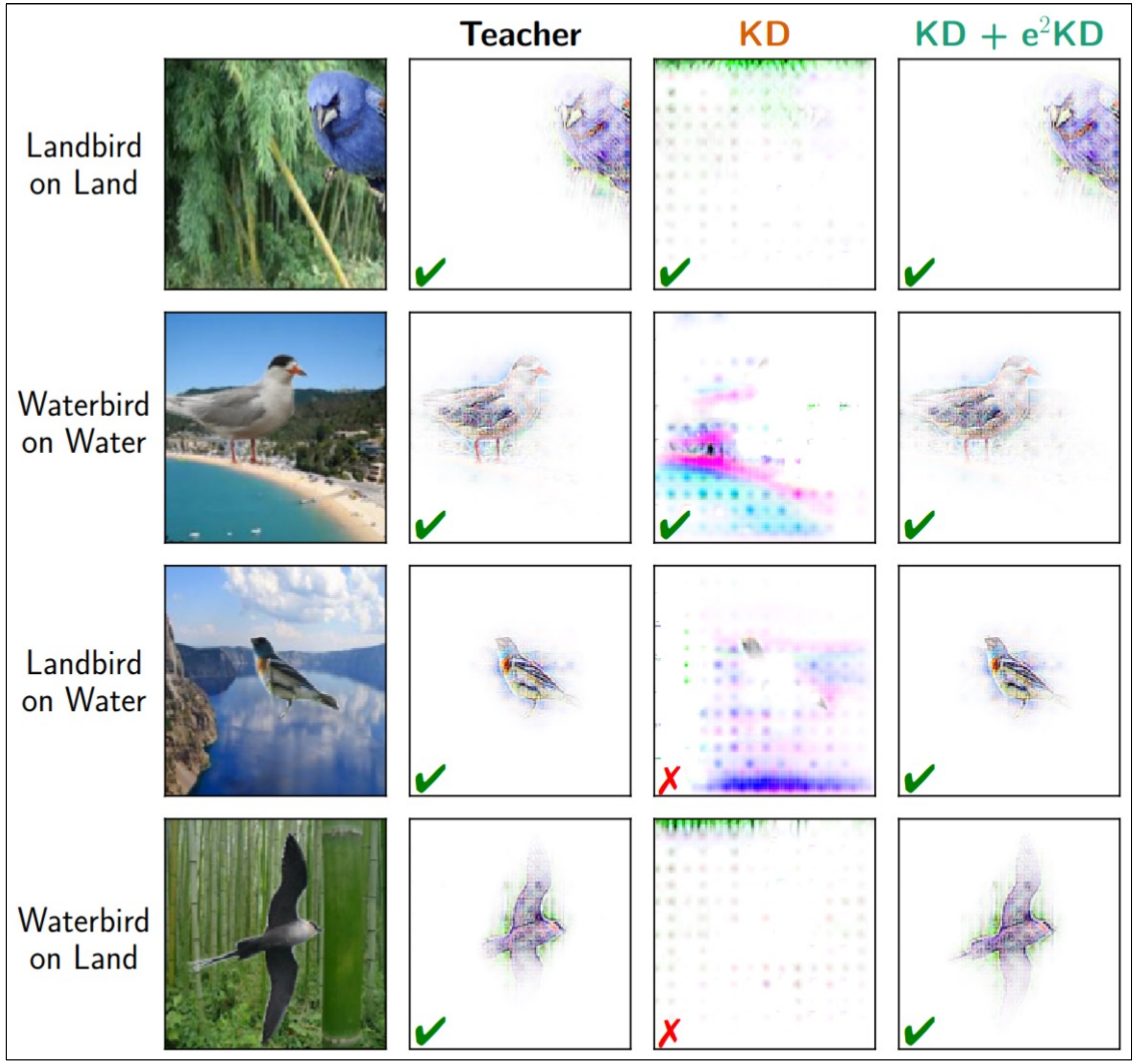
Desideratum 2: Learning the 'Right' Features

Task: Classify Landbird vs. Waterbird

Distillation Data: Landbird on Land, Waterbird on Water

Test-time: Landbird on Water, Waterbird on Land

Teacher: ResNet-50 explicitly guided to focus on the bird



Focusing on the 'Right' input features gives OOD robustness.

The student might deviate from the teacher!

e²KD effectively maintains correct reasoning!

Desideratum 3: Maintaining Interpretability

Distill a teacher with desirable explanations!

1. **Due to its training**
2. **Due to its architecture**

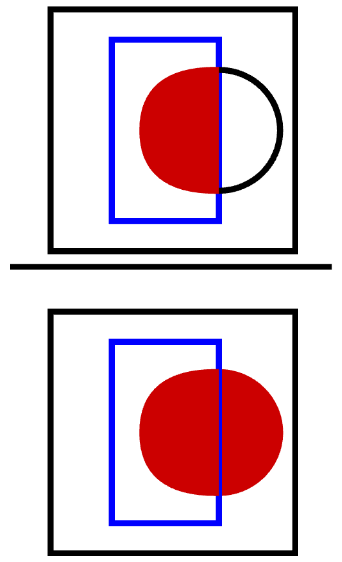
Desideratum 3: Maintaining Interpretability

Distill a teacher with desirable explanations!

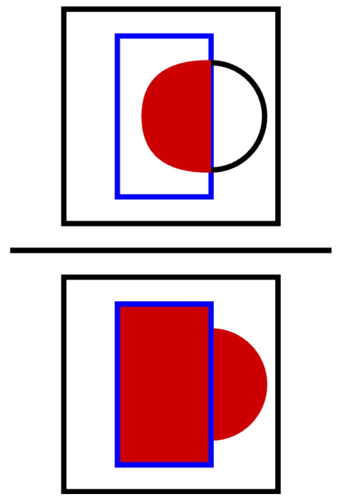
1. **Due to its training**

   Pascal VOC as multi-label classification

EPG Score:

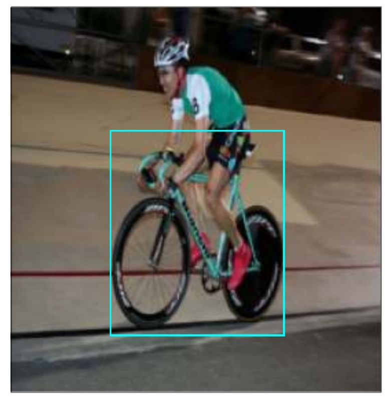


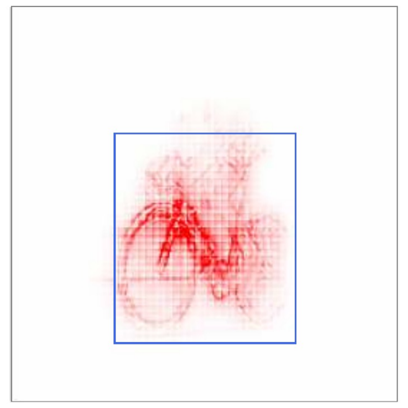
IoU Score:



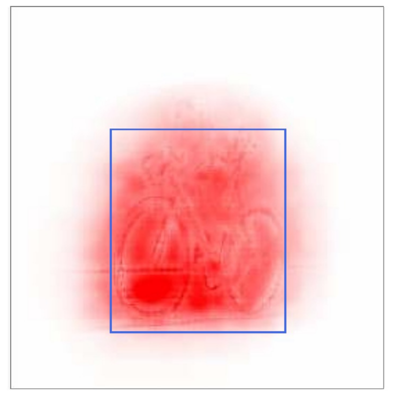
	EPG Teacher			IoU Teacher		
	EPG	IoU	F1	EPG	**IoU**	F1
Teacher ResNet-50	75.7	21.3	72.5	65.0	49.7	72.8

bicycle

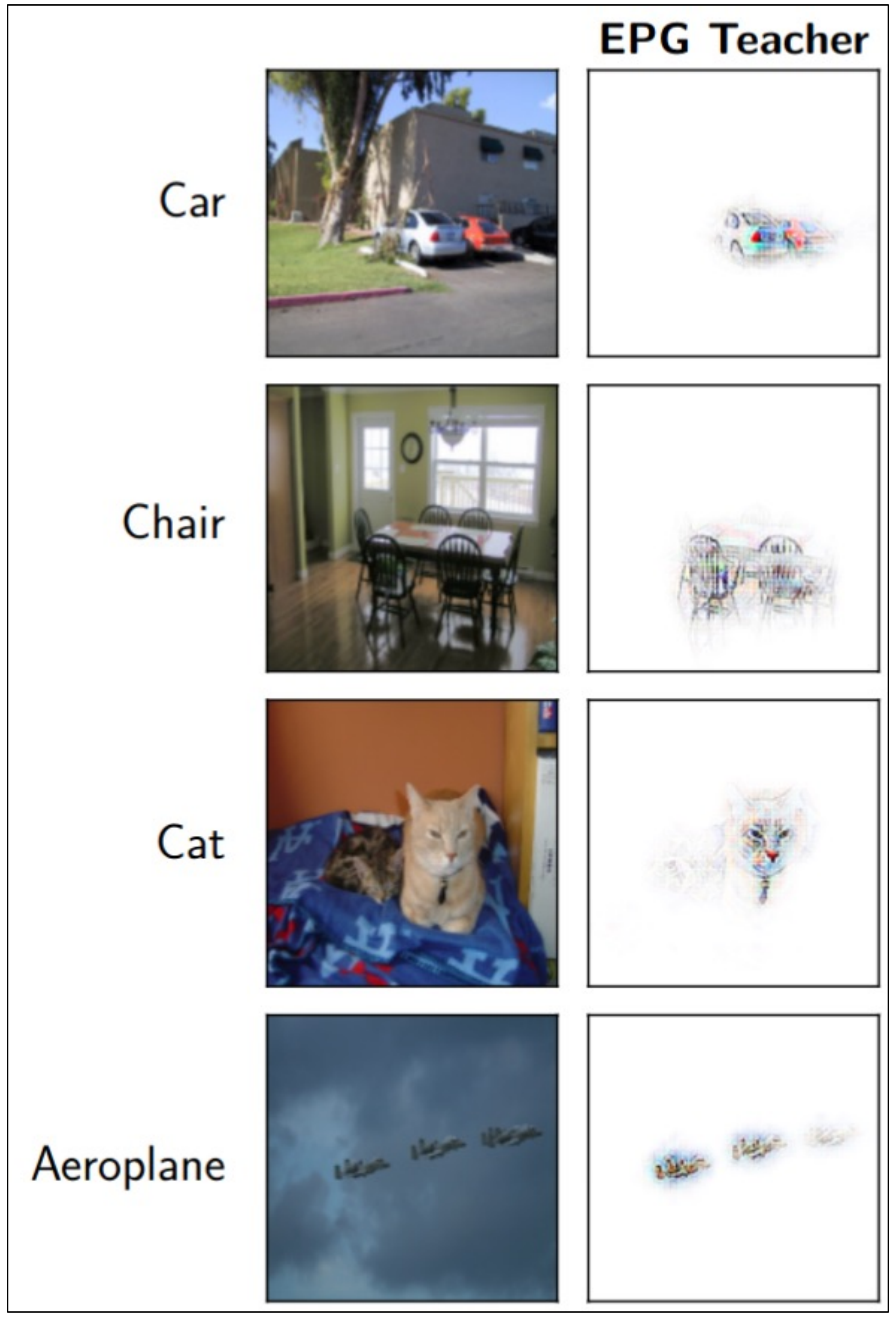




High EPG



High IoU



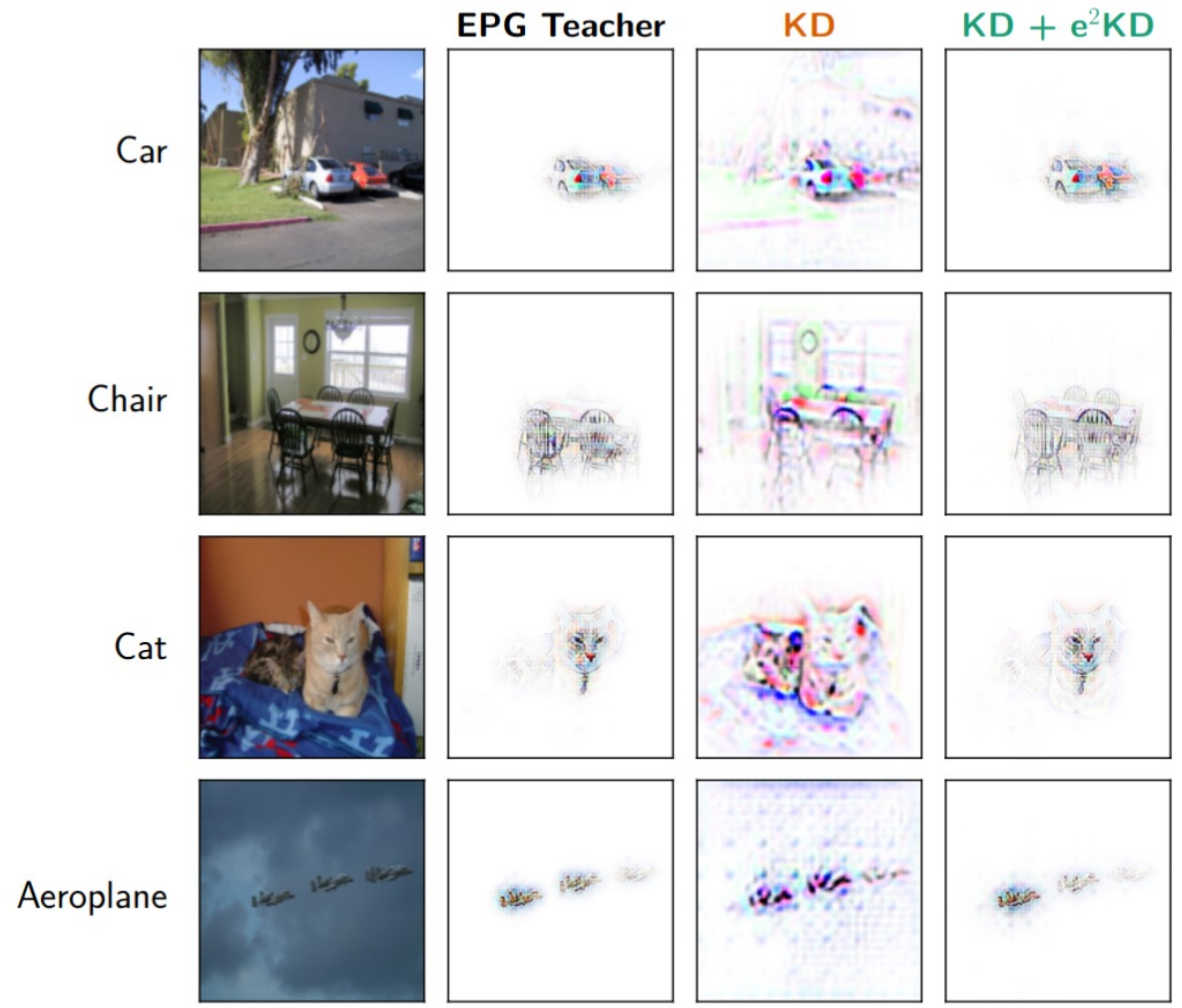
Desideratum 3: Maintaining Interpretability

Distill a teacher with desirable explanations!

1. **Due to its training**

   Pascal VOC as multi-label classification

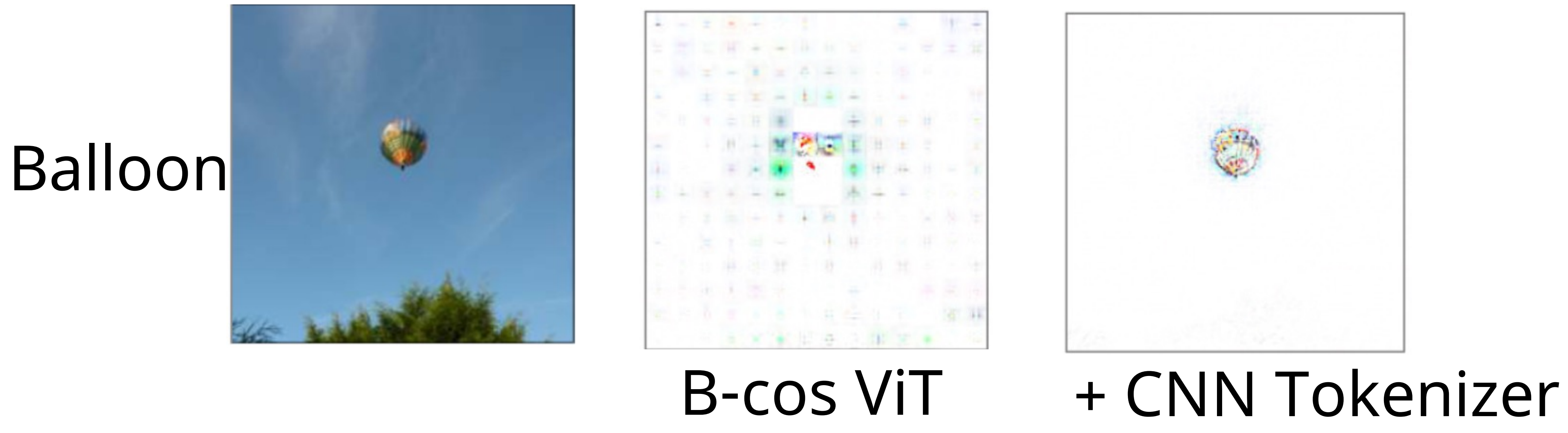
	EPG Teacher			IoU Teacher		
	EPG	IoU	F1	EPG	**IoU**	F1
Teacher ResNet-50	75.7	21.3	72.5	65.0	49.7	72.8
Baseline ResNet-18	50.0	29.0	58.0	50.0	29.0	58.0
KD [38]	60.1	31.6	60.1	58.9	35.7	62.7
+ e^2KD (B-cos)	**71.1**	24.8	**67.6**	60.3	**45.7**	**64.8**

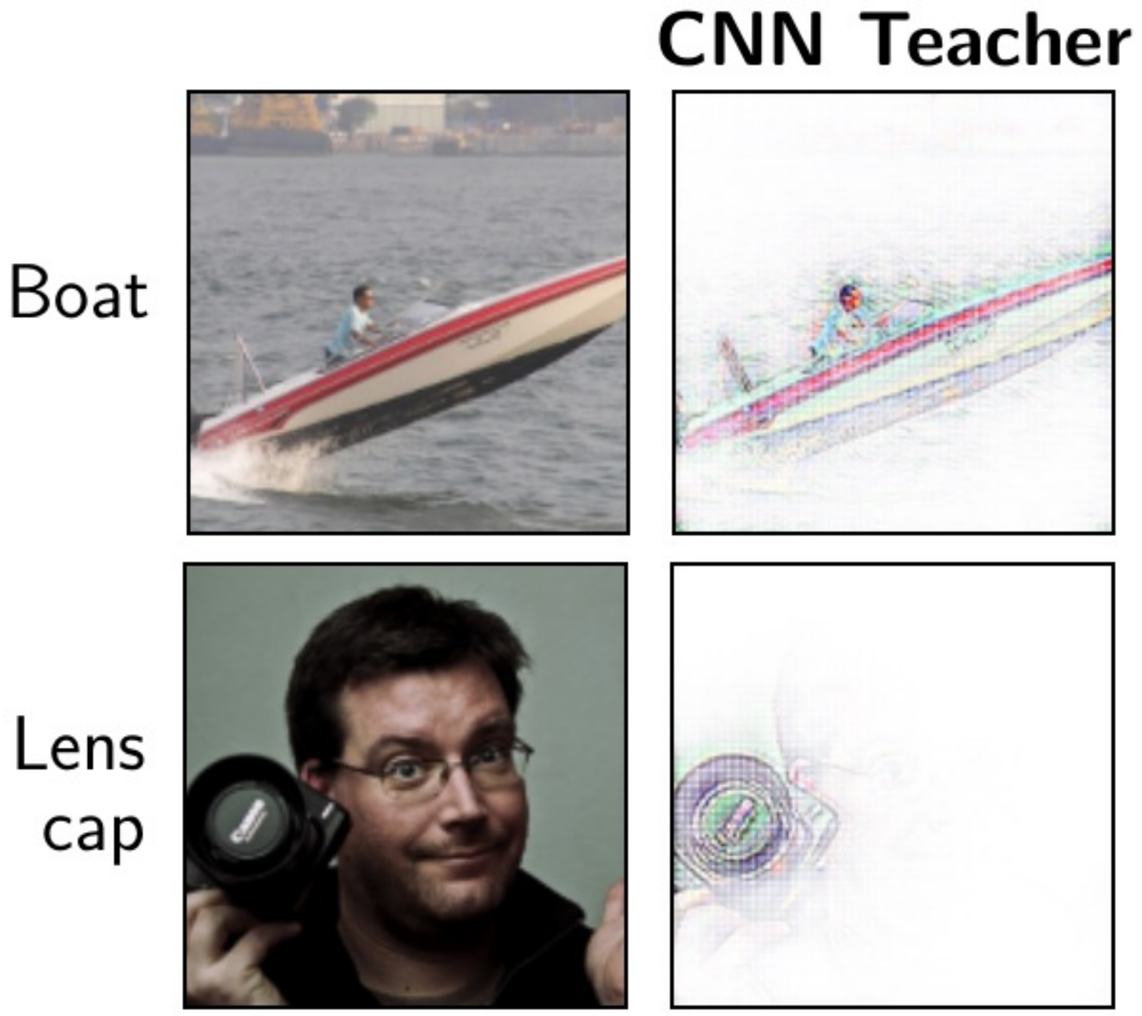


Desideratum 3: Maintaining Interpretability

Distill a teacher with desirable explanations!

1. Due to its training
2. **Due to its architecture**





Can we instead *distill* such a prior?

Distillation: B-cos DenseNet-169 → B-cos ViT_{Tiny}
Setting: ImageNet

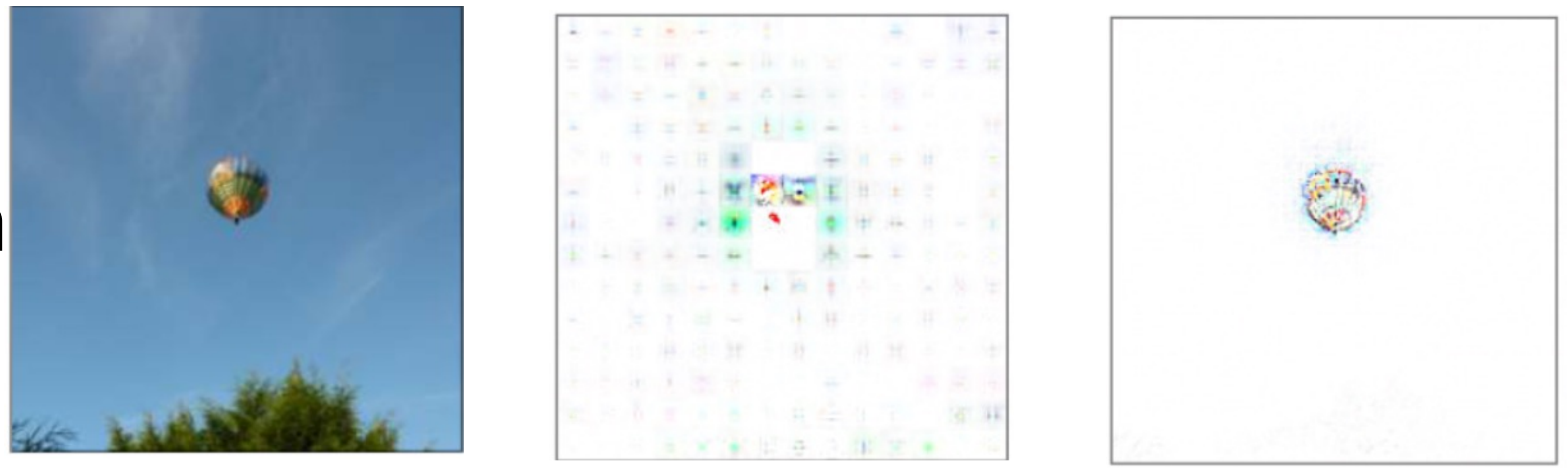
Böhle et al., B-cos Alignment for Inherently Interpretable CNNs and Vision Transformers, TPAMI 2024

Desideratum 3: Maintaining Interpretability

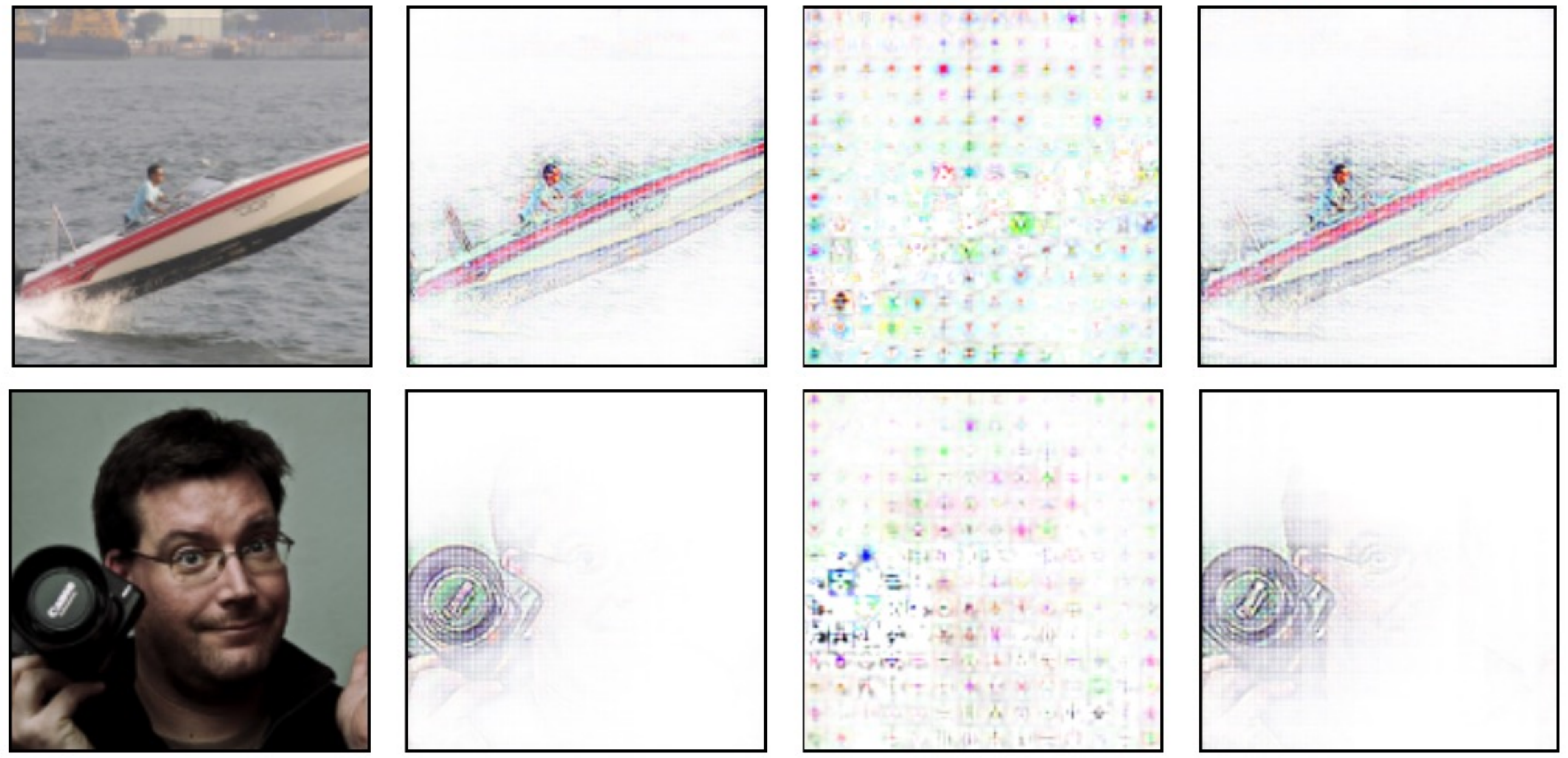
Distill a teacher with desirable explanations!

1. Due to its training
2. **Due to its architecture**

Balloon



B-cos ViT   + CNN Tokenizer



Can we instead *distill* such a prior?

Distillation: B-cos DenseNet-169 → B-cos ViT_{Tiny}
Setting: ImageNet

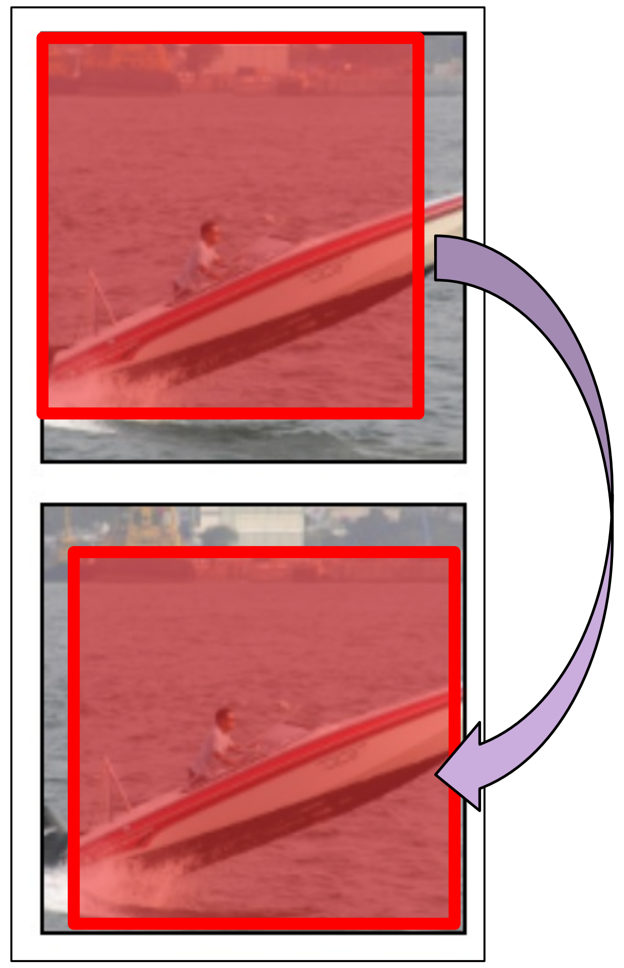
Method	Acc.	Agr.
T: B-cos DenseNet-169	75.2	-
B: B-cos ViT_{Tiny}	60.0	64.6
KD	64.8	70.1
+ e^2KD	**66.3**	**71.8**

Böhle et al., B-cos Alignment for Inherently Interpretable CNNs and Vision Transformers, TPAMI 2024

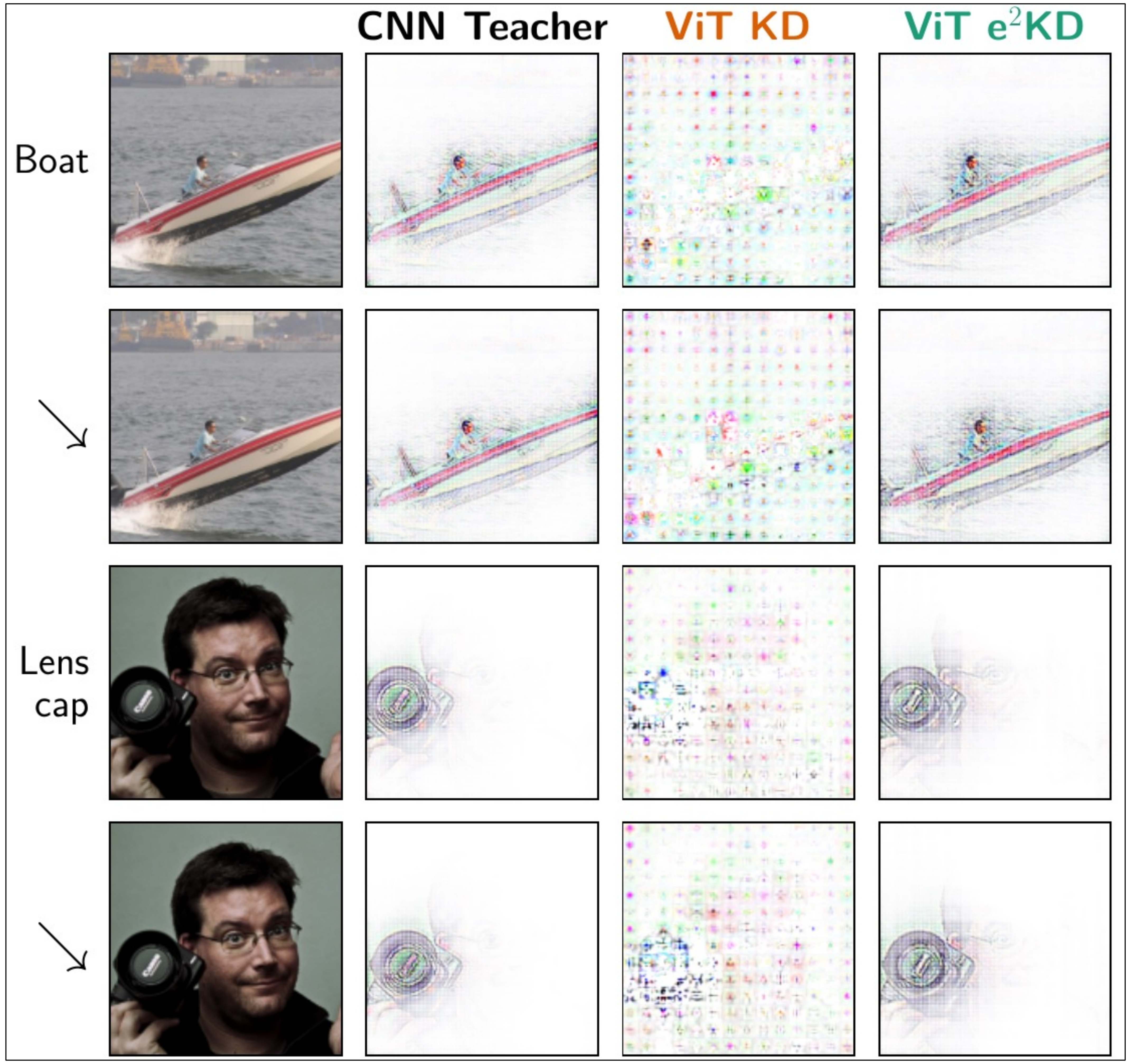
Desideratum 3: Maintaining Interpretability

Distill a teacher with desirable explanations!

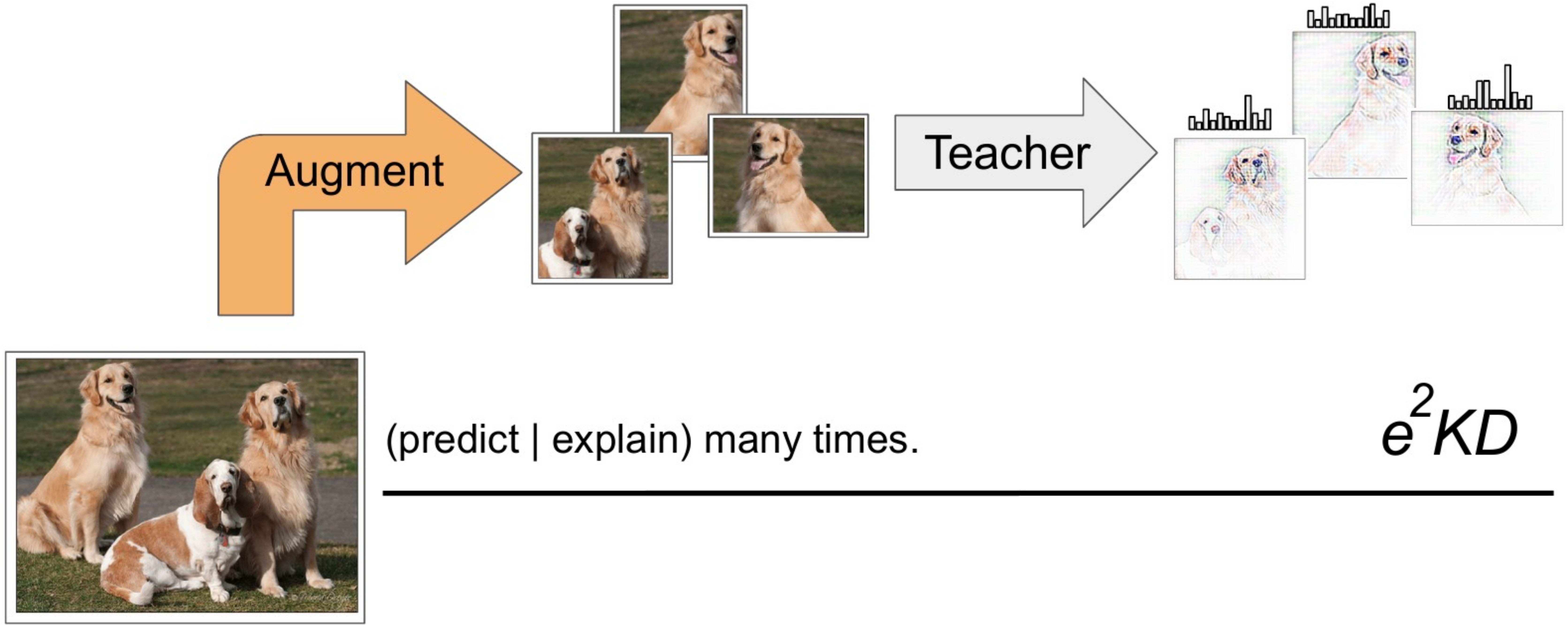
1. Due to its training
2. **Due to its architecture**



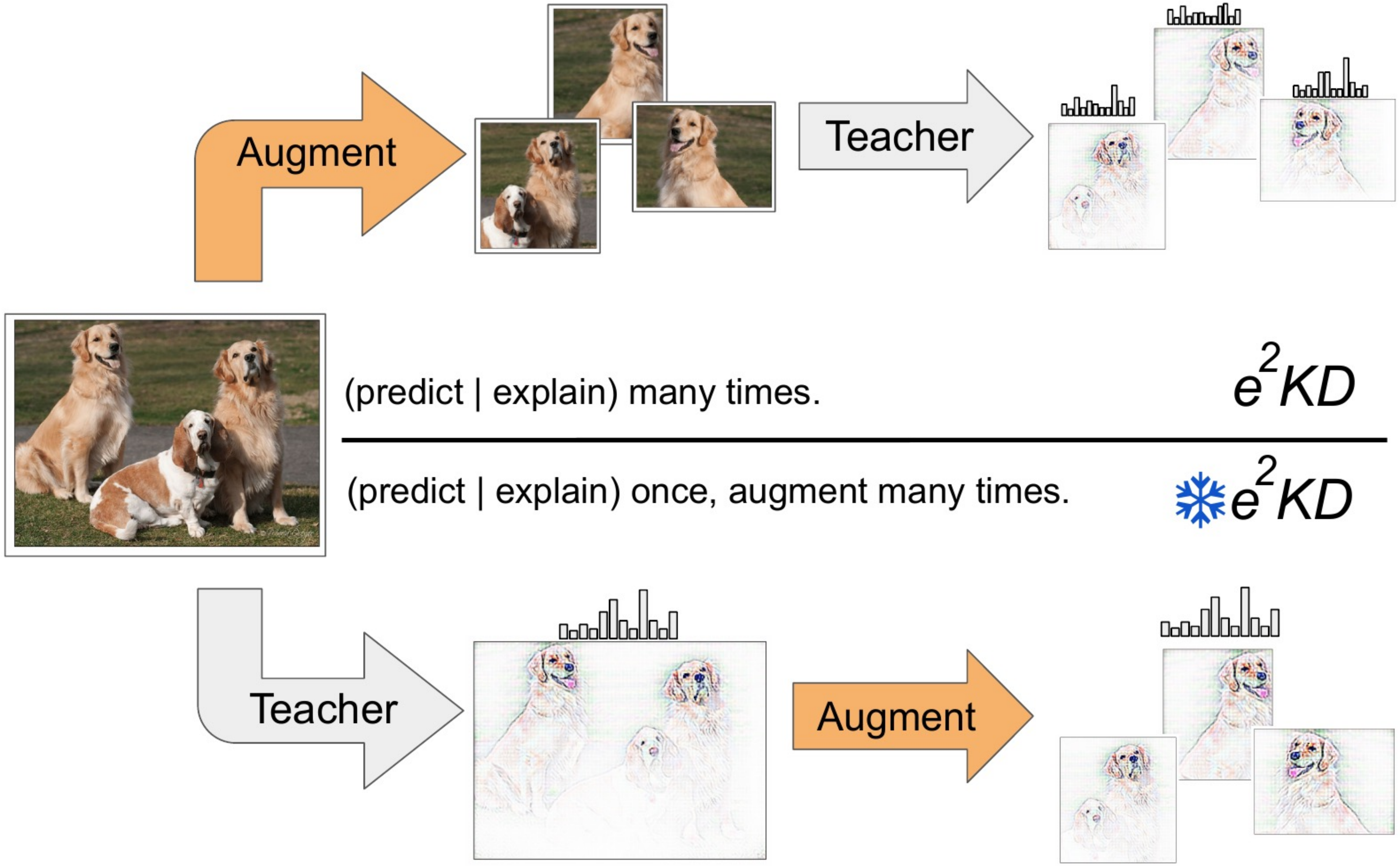
Measuring shift-equivariance



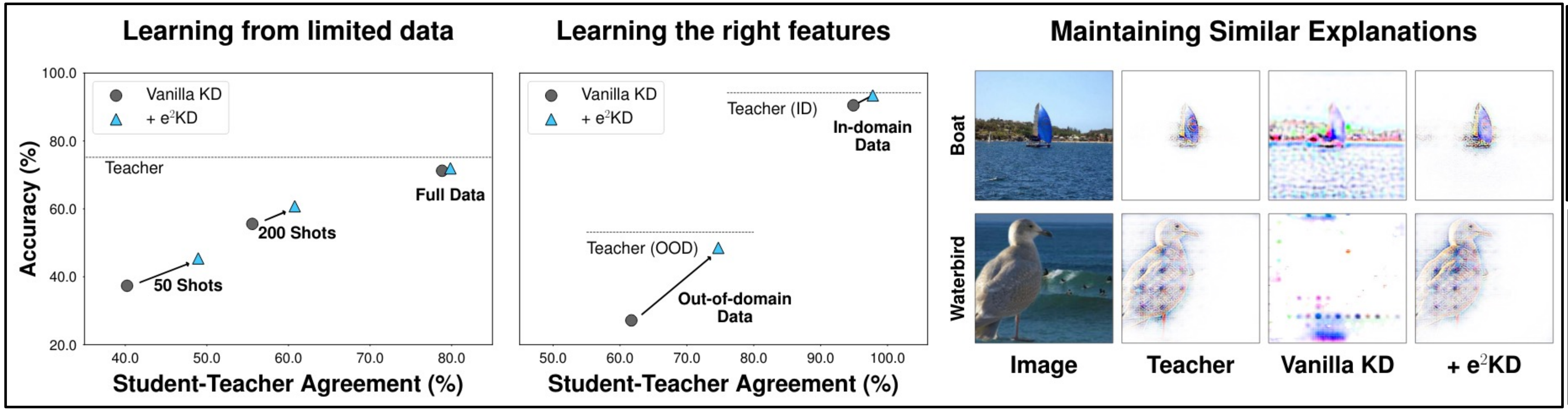
e²KD with Frozen Explanations

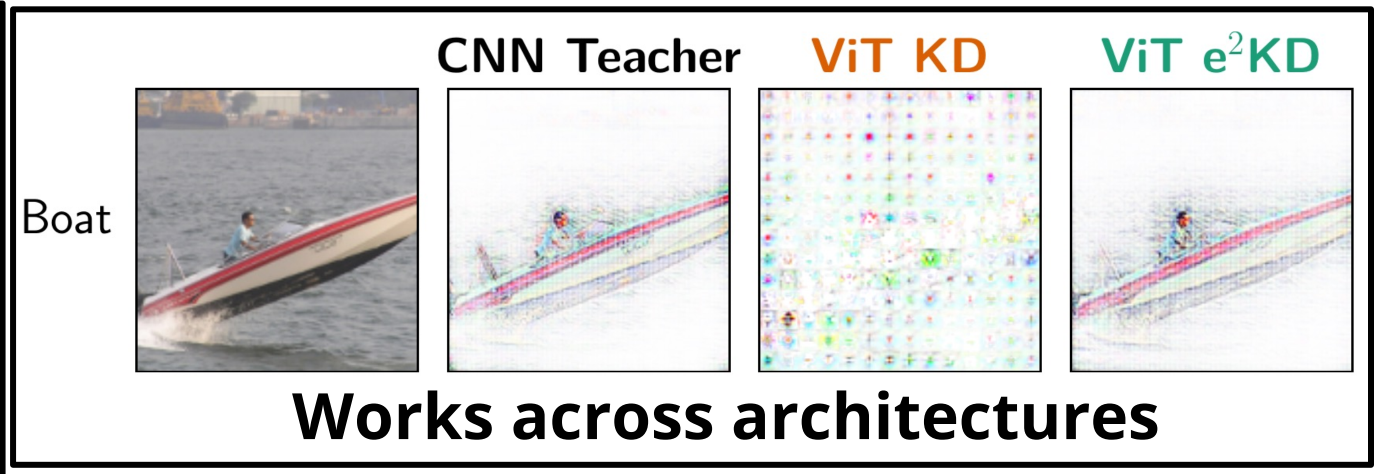


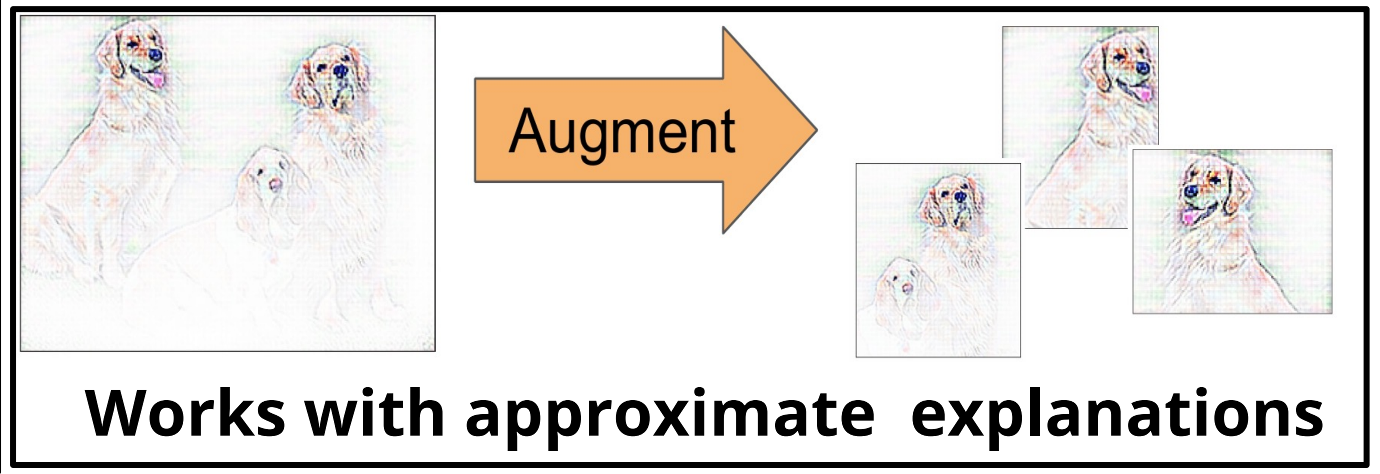
e^2KD with Frozen Explanations



Good Teachers Explain: Explanation-enhanced Knowledge Distillation







Poster ID: #330

Poster Session: Tue 1 Oct 2024, 10:30 a.m. — 12:30 p.m. CEST

Paper
https://arxiv.org/abs/2402.03119

Code
github.com/m-parchami/GoodTeachersExplain

Contact
mparcham@mpi-inf.mpg.de