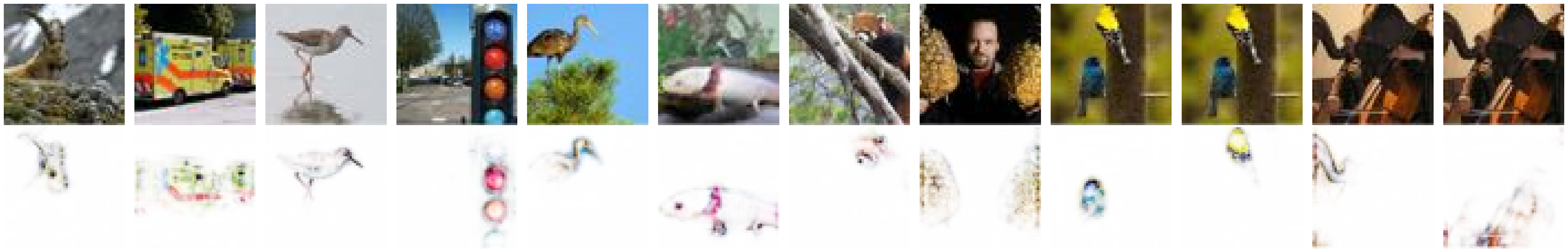
Interpretability for Deep Learning in Computer Vision


Moritz Boehle


Sukrut Rao


Mario Fritz
CISPA Helmholtz


Bernt Schiele

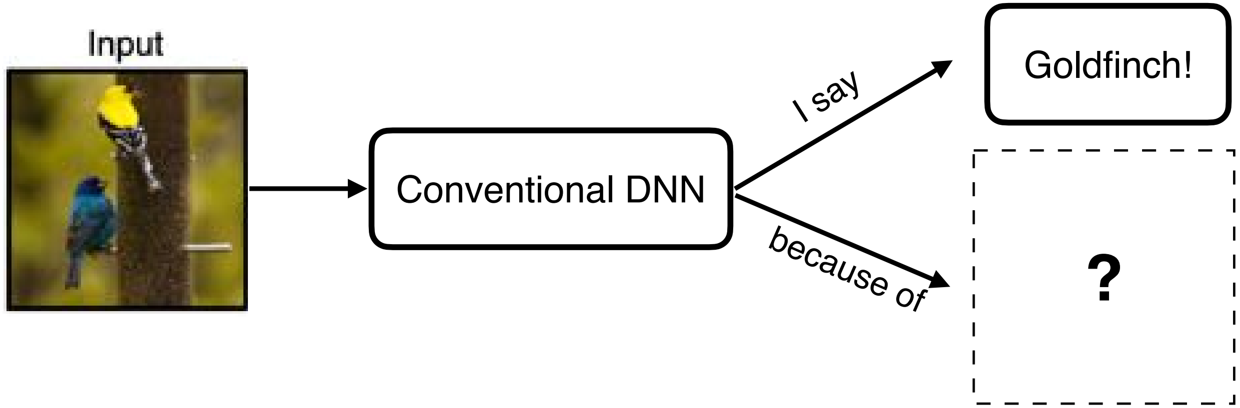


Max Planck Institute for Informatics & Saarland University,
Saarland Informatics Campus Saarbrücken

Motivation

Want Explanation to understand better / detect bias / gain trust

Requirements Faithful Interpretable

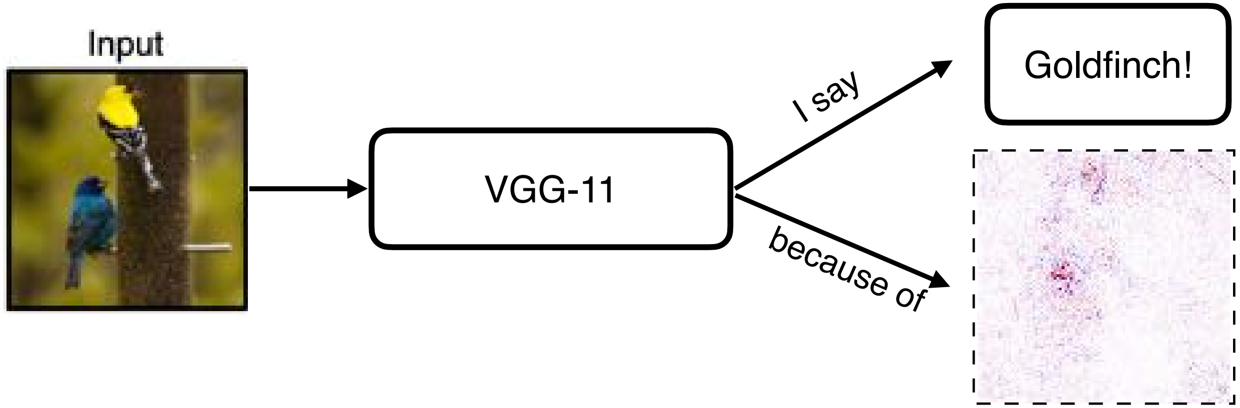


References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

Want Explanation

Requirements Faithful ✓ Interpretable ✗



Gradient

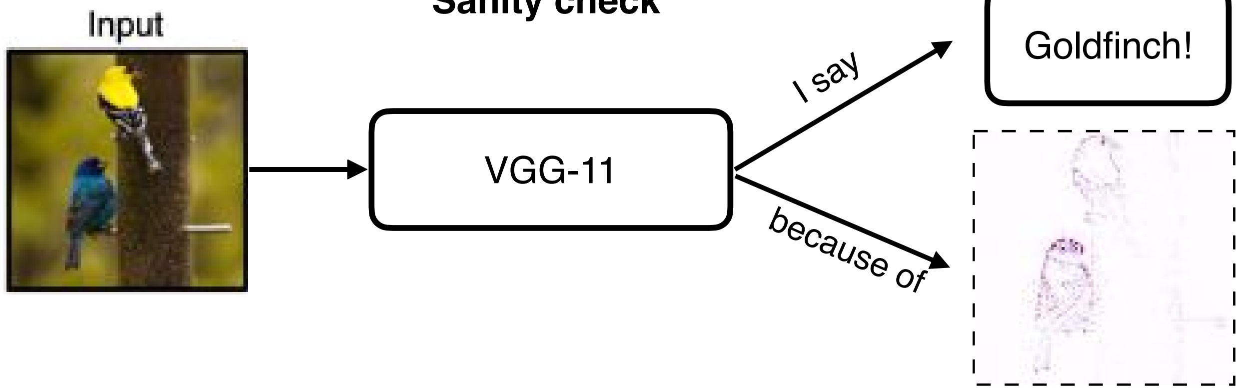
References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

Want | Explanation

Requirements | Faithful | Interpretable ✓

Sanity check



Guided Backpropagation

References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

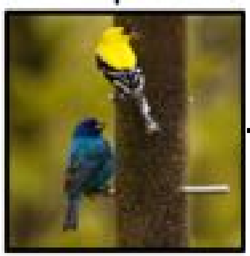
Motivation

Want — Explanation

Requirements — Faithful — Interpretable **?**

Sanity check

Input

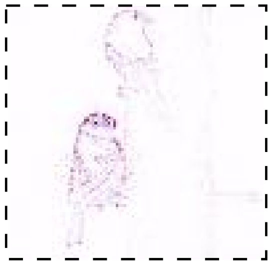


Randomised 1 layer

VGG-11

I say → Goldfinch!

because of →



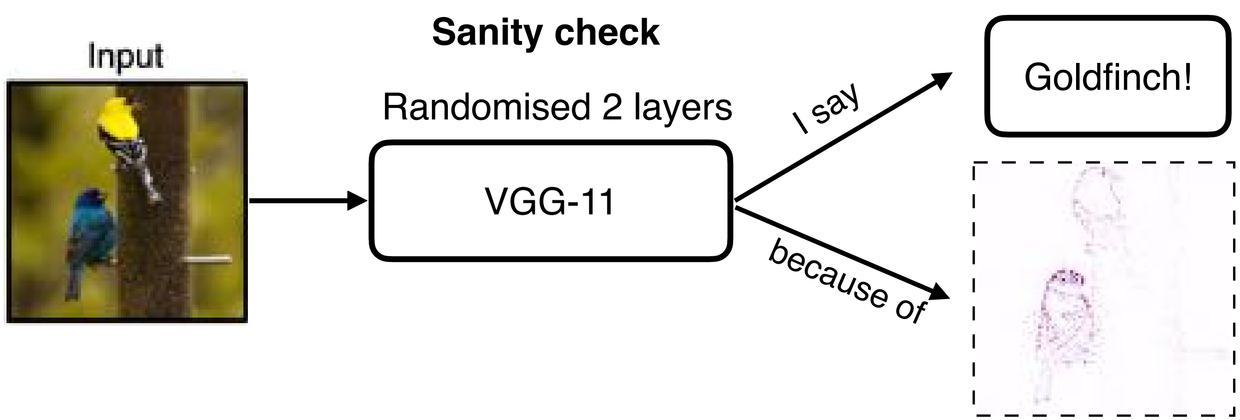
Guided Backpropagation

References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

Want: Explanation

Requirements: Faithful | Interpretable ?



Guided Backpropagation

References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

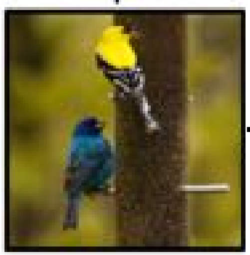
Motivation

Want | Explanation

Requirements | Faithful | Interpretable ?

Sanity check

Input

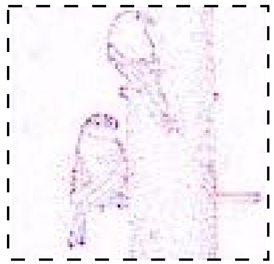


Randomised 5 layers

VGG-11

I say → Goldfinch!

because of →



Guided Backpropagation

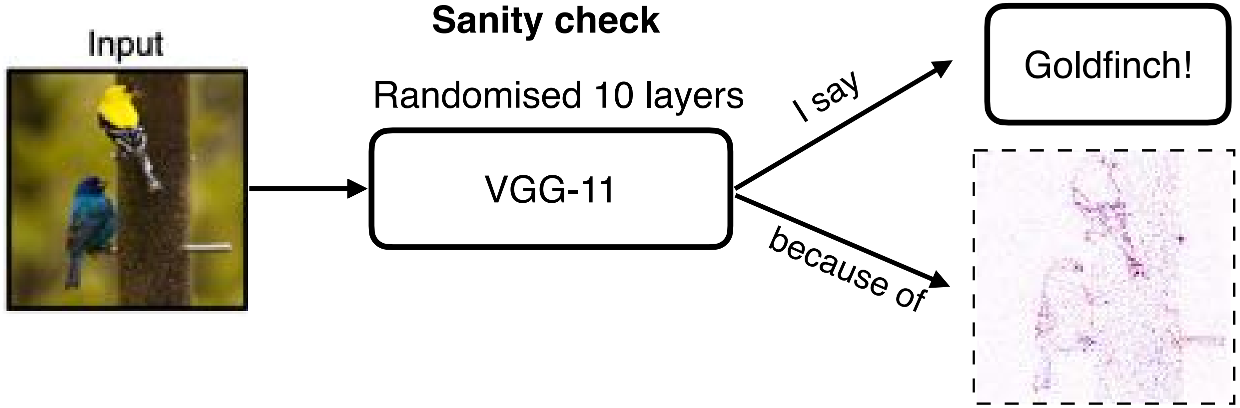
References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

Want | Explanation

Requirements | Faithful ✗ | Interpretable ✓

Sanity check



Guided Backpropagation

References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Overview

- **Interpretability** for Deep Learning in Computer Vision
  - Towards Better **Understanding** of **Attribution** Methods — CVPR'22, arXiv'23 [2303.11884]
  - **Inherently Interpretable** CNN Networks — CVPR'21, CVPR'22
  - **Inherently Interpretable** Transformer Networks — arXiv'23 [2301.08571]
  - Using **Explanations** to **Guide Inherently Interpretable** Models — ICCV'23

Towards Better Understanding of Attribution Methods

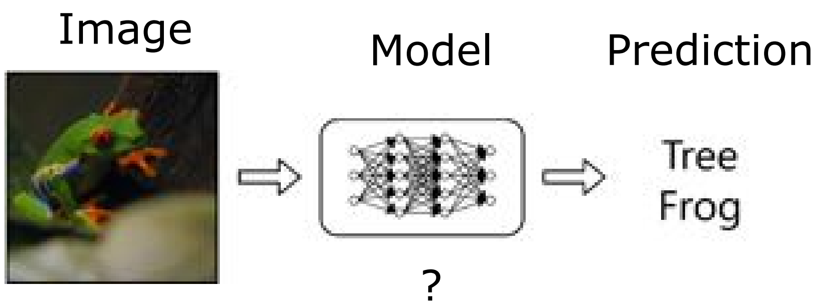
@ CVPR 2022 — extended version @ arXiv 2023

Sukrut Rao
MPI Informatics

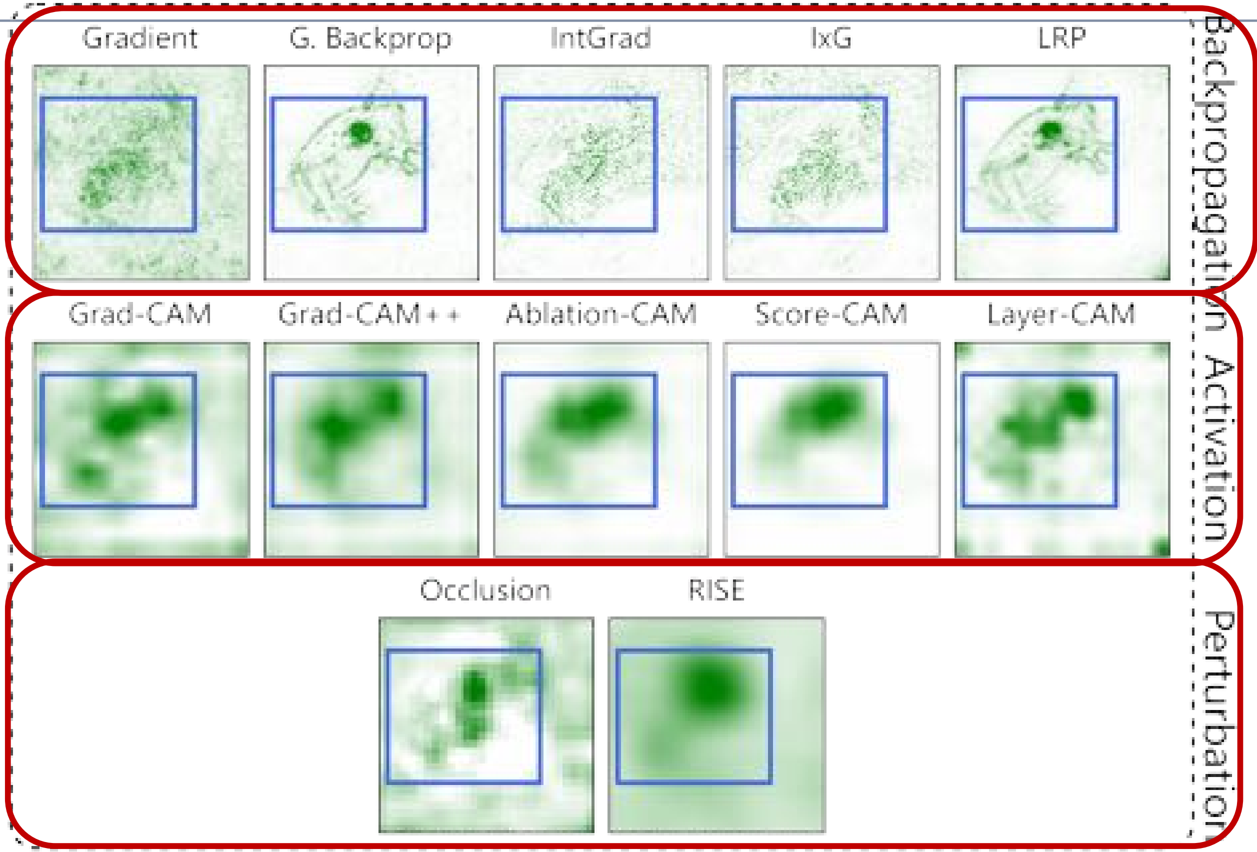
Moritz Boehle
MPI Informatics

Bernt Schiele
MPI Informatics

(Post-Hoc) Attribution Methods

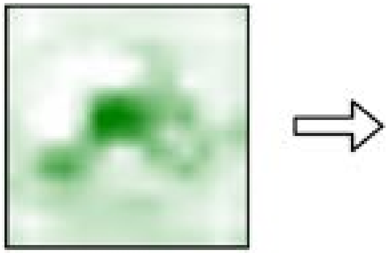


- Major Issues
  - they are surprisingly different !
    not all can be faithful !
  - evaluation challenging, as
    ground truth attributions unknown



Gradient (Simonyan et al., 2014), G. Backprop (Springenberg et al., 2015), IntGrad (Sundararajan et al., 2017), IxG (Shrikumar et al., 2017), Grad-CAM (Selvaraju et al., 2017), Grad-CAM++ (Chattopadhyay et al., 2018), Ablation-CAM (Desai et al., 2020), Score-CAM (Wang et al., 2020), Layer-CAM (Jiang et al., 2021), Occlusion (Zeiler et al., 2014), RISE (Petsiuk et al., 2018)

Evaluating Attribution Methods: Object Localization

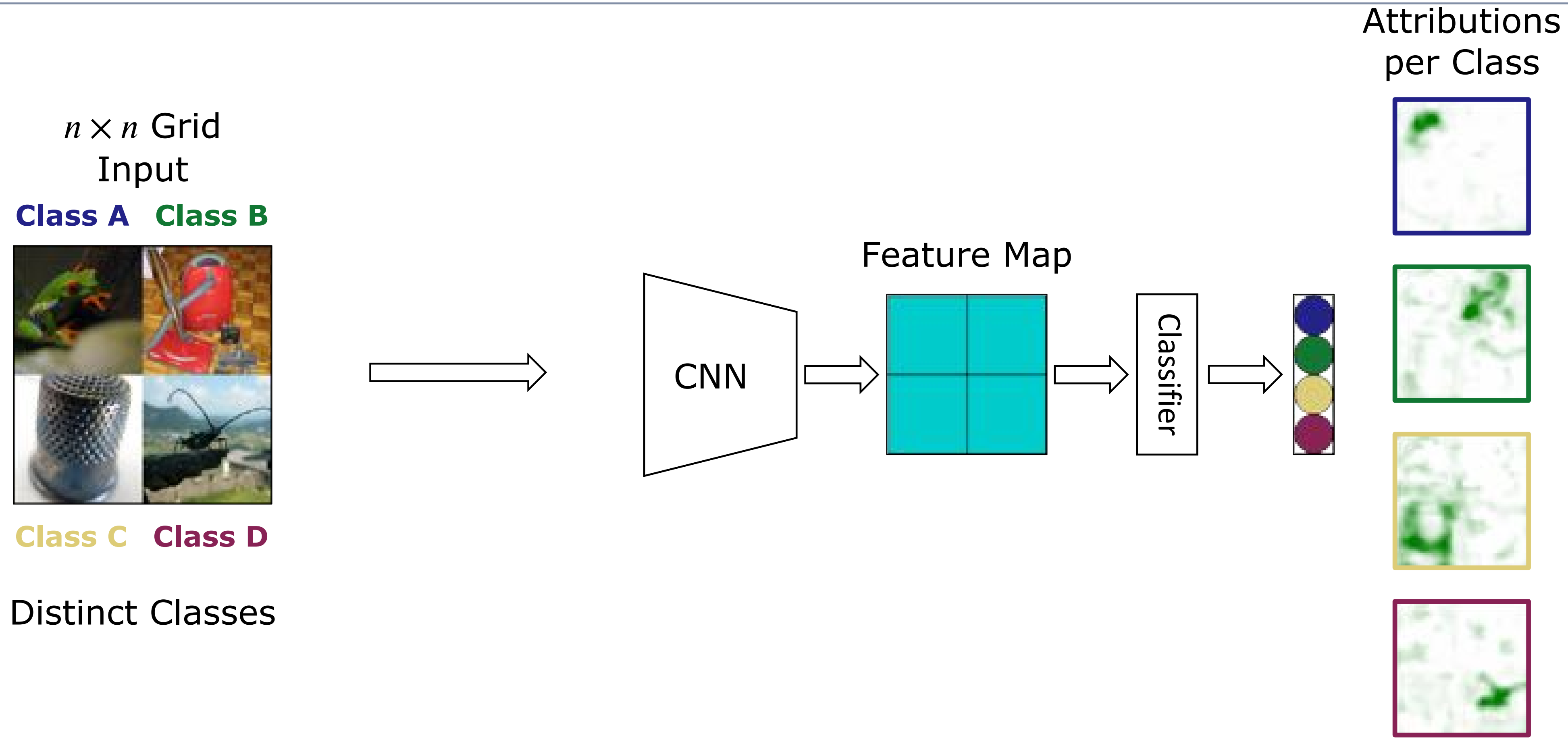


Attributions

Object
Bounding Box

Cao et al. Look and Think Twice: Capturing Top-Down Visual Attention with Feedback Convolutional Neural Networks. ICCV 2015.

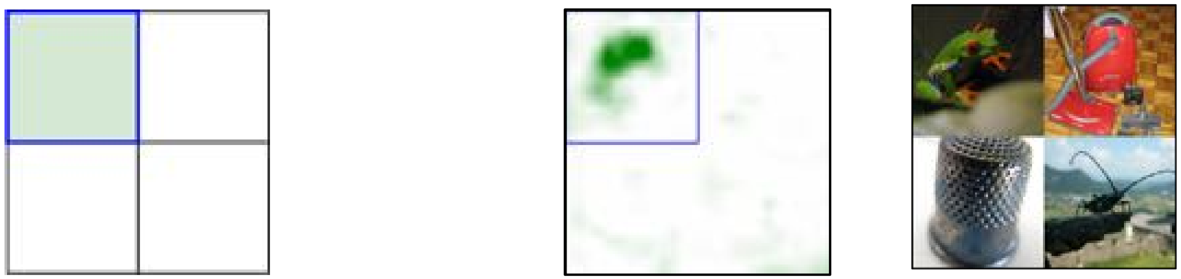
Evaluating Attribution Methods: Grid Pointing Game (GridPG)



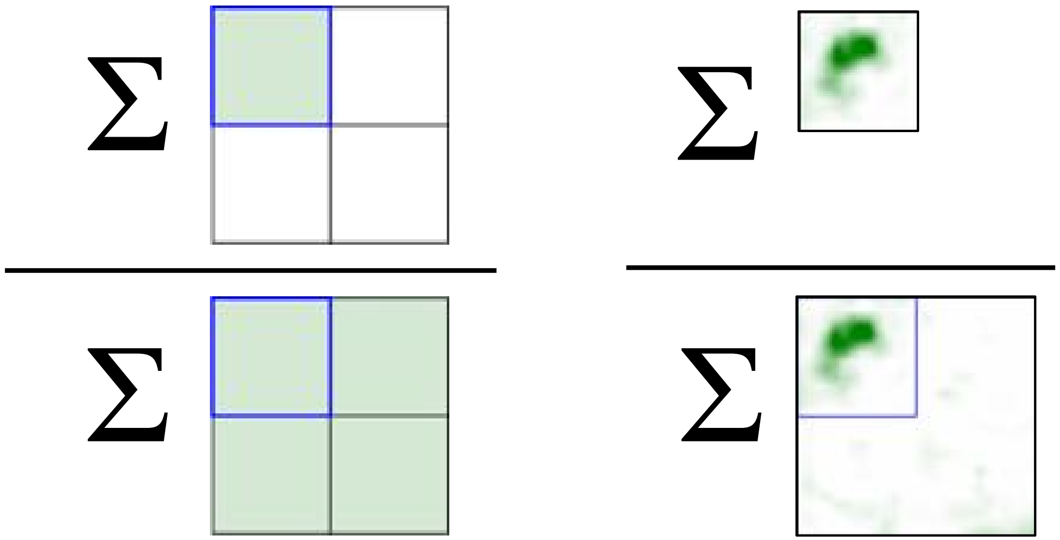
Moritz Böhle, Mario Fritz, Bernt Schiele. Convolutional Dynamic Alignment Networks for Interpretable Classifications. CVPR 2021.

Evaluating Attribution Methods: Grid Pointing Game (GridPG)

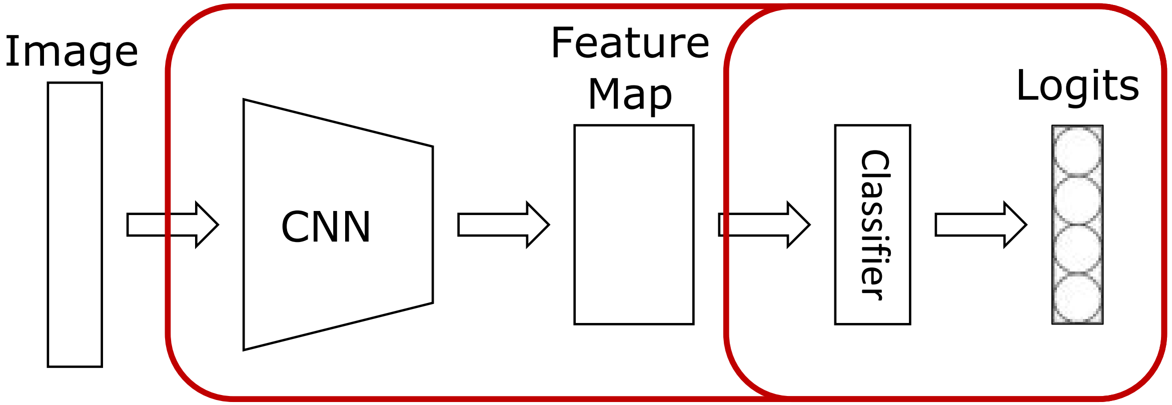
- Expected Localization:
  (for top-left)

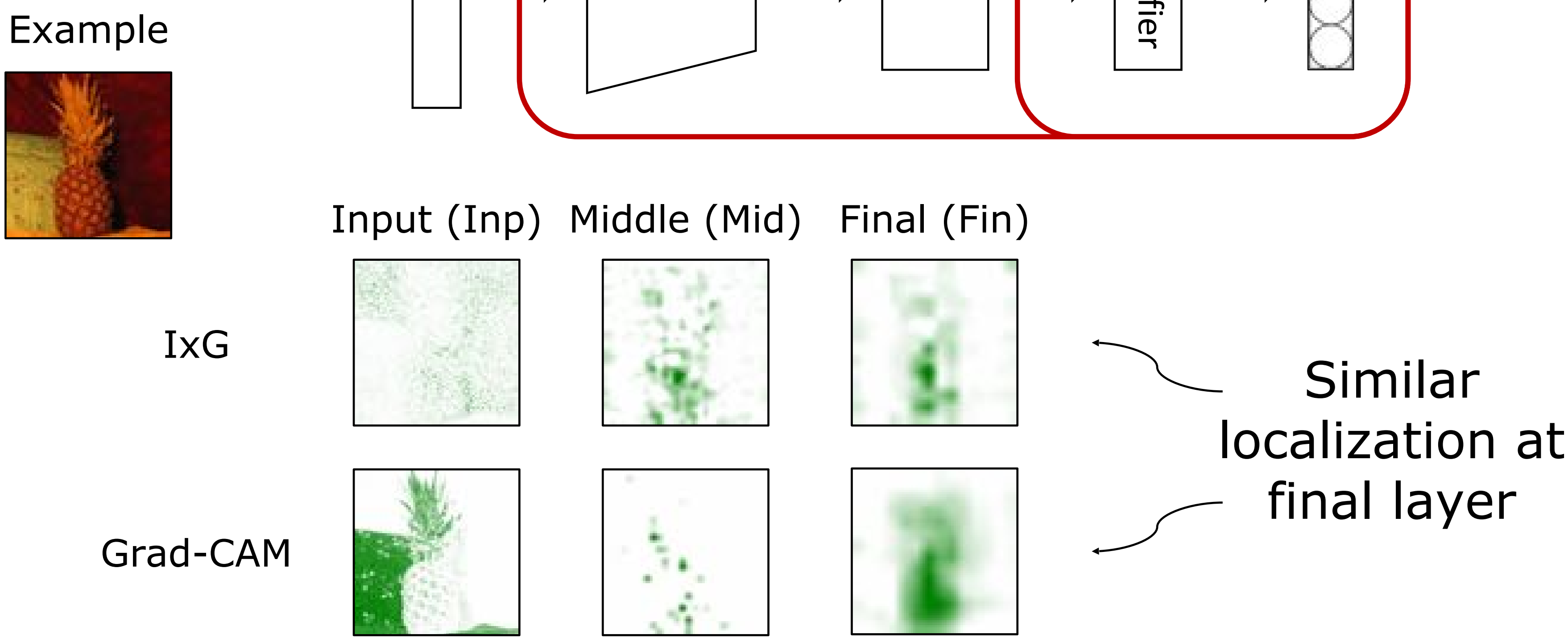


- Localization Metric:



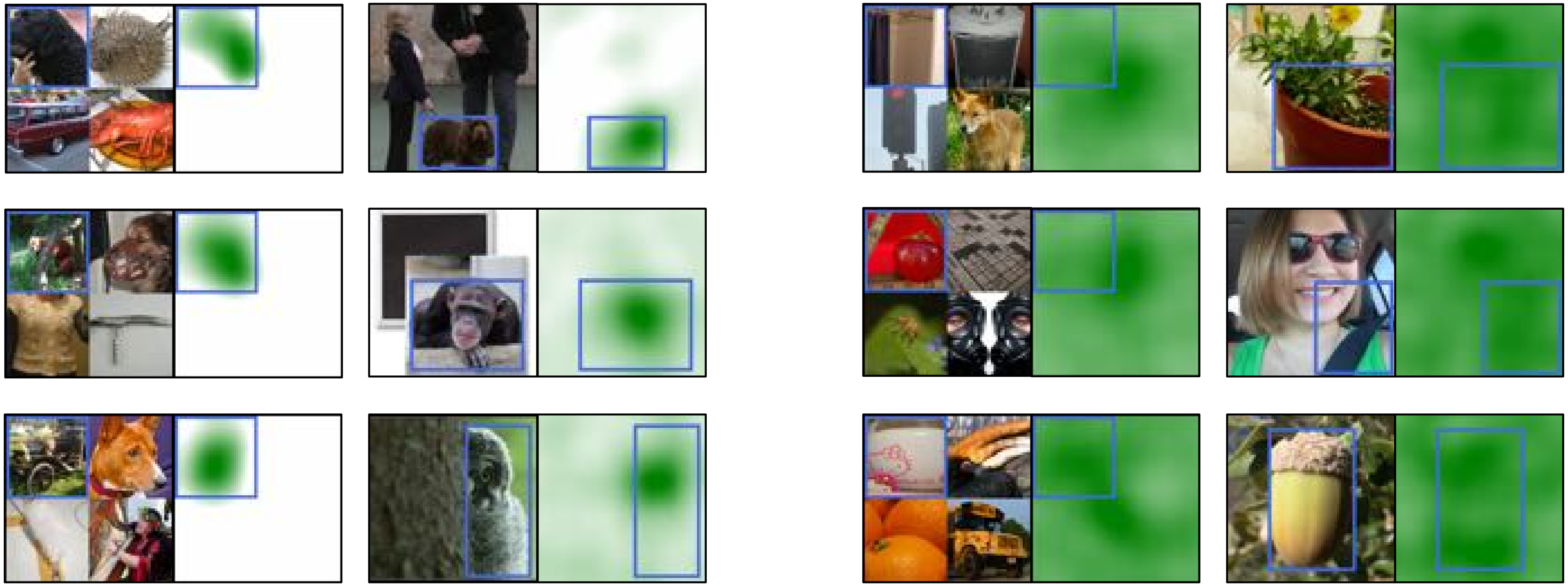
Multi-Layer Attribution Evaluation: ML-Att





Qualitative Evaluation

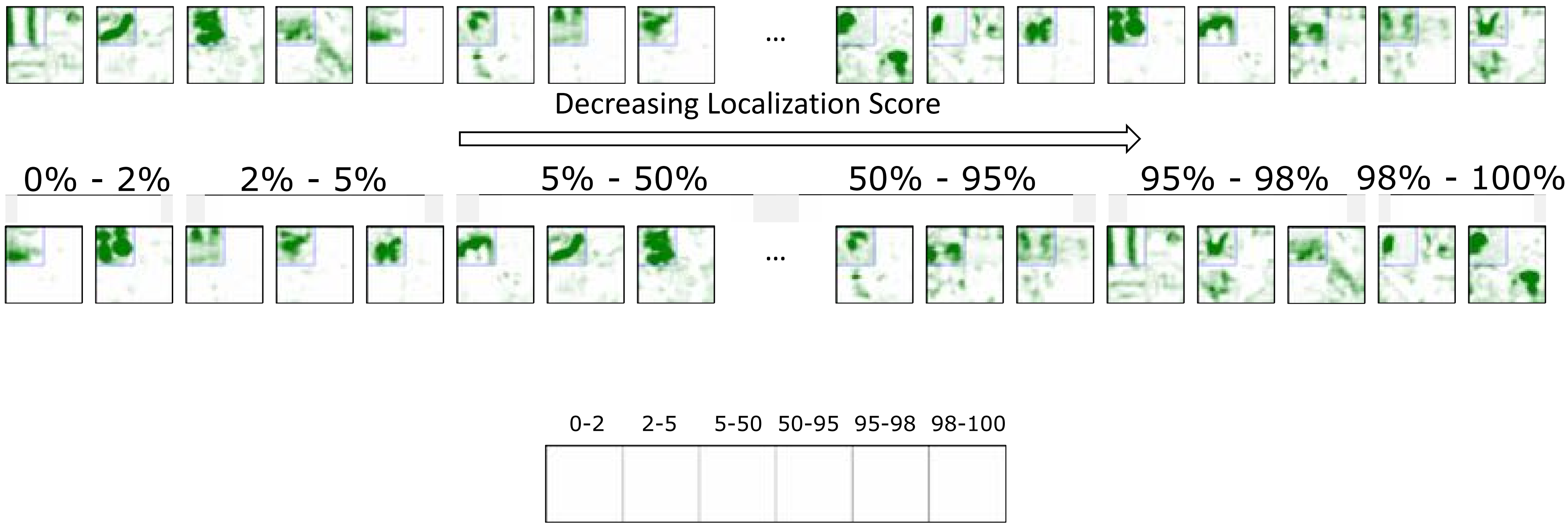
Well Localized | Poorly Localized



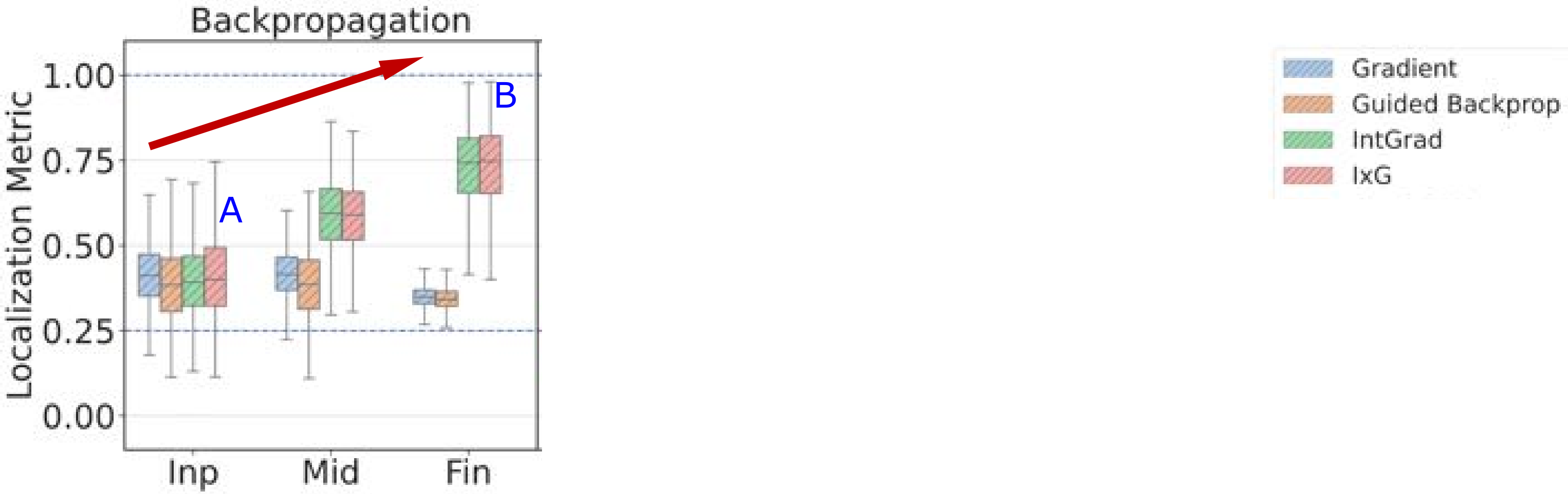
Example: RISE

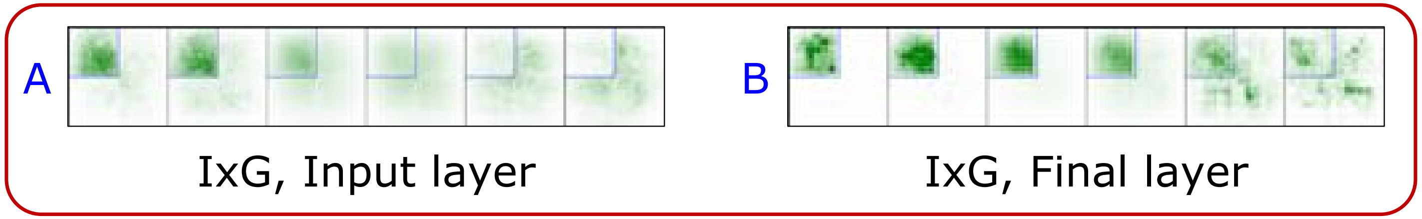
Systematic Qualitative Evaluation: Aggregate Attribution Evaluation: AggAtt

1. Sort
2. Bin
3. Aggregate

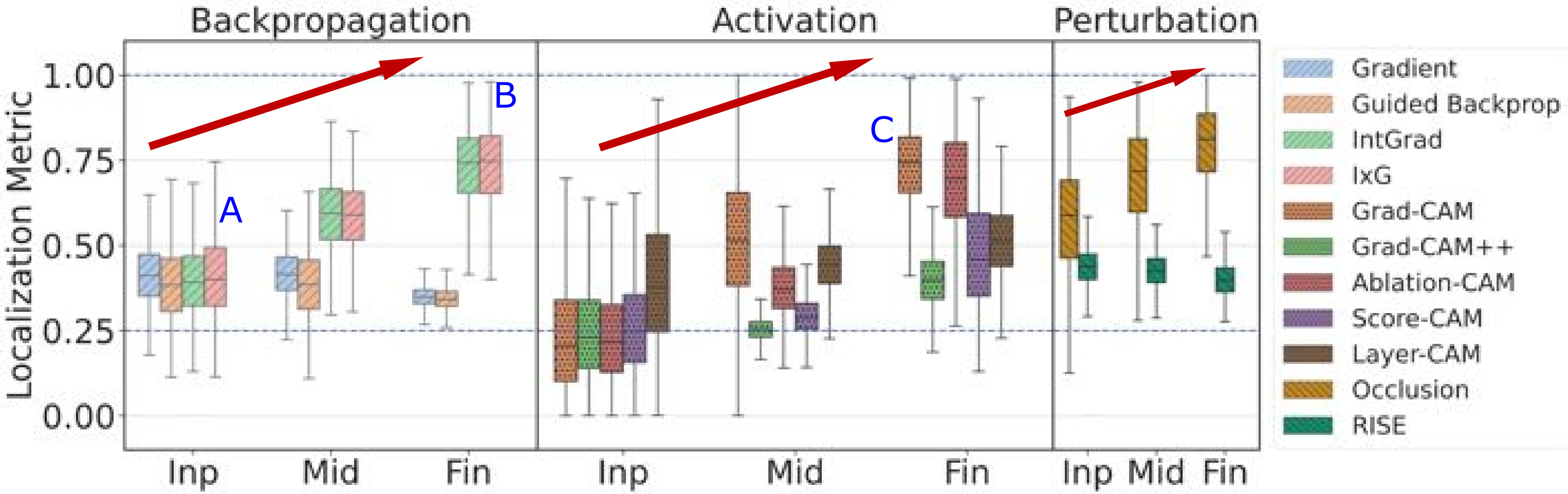


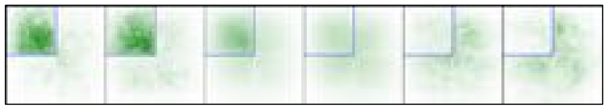
Results: GridPG



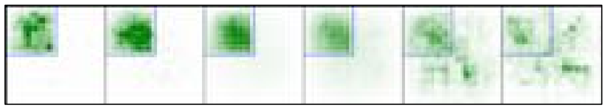


Results: GridPG

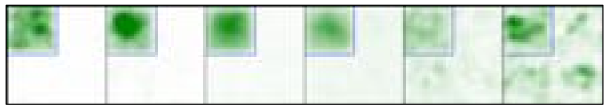


A 

IxG, Input layer

B 

IxG, Final layer

C 

Grad-CAM, Final layer

Similar localization at final layer

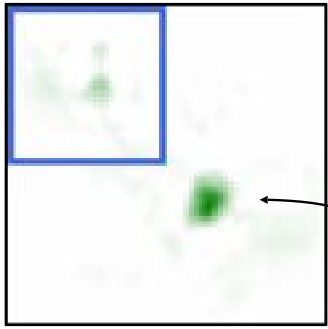
Grid Pointing Game (GridPG)

Challenge: Classification of each cell possibly influenced by others

Class A Class B

Class C Class D

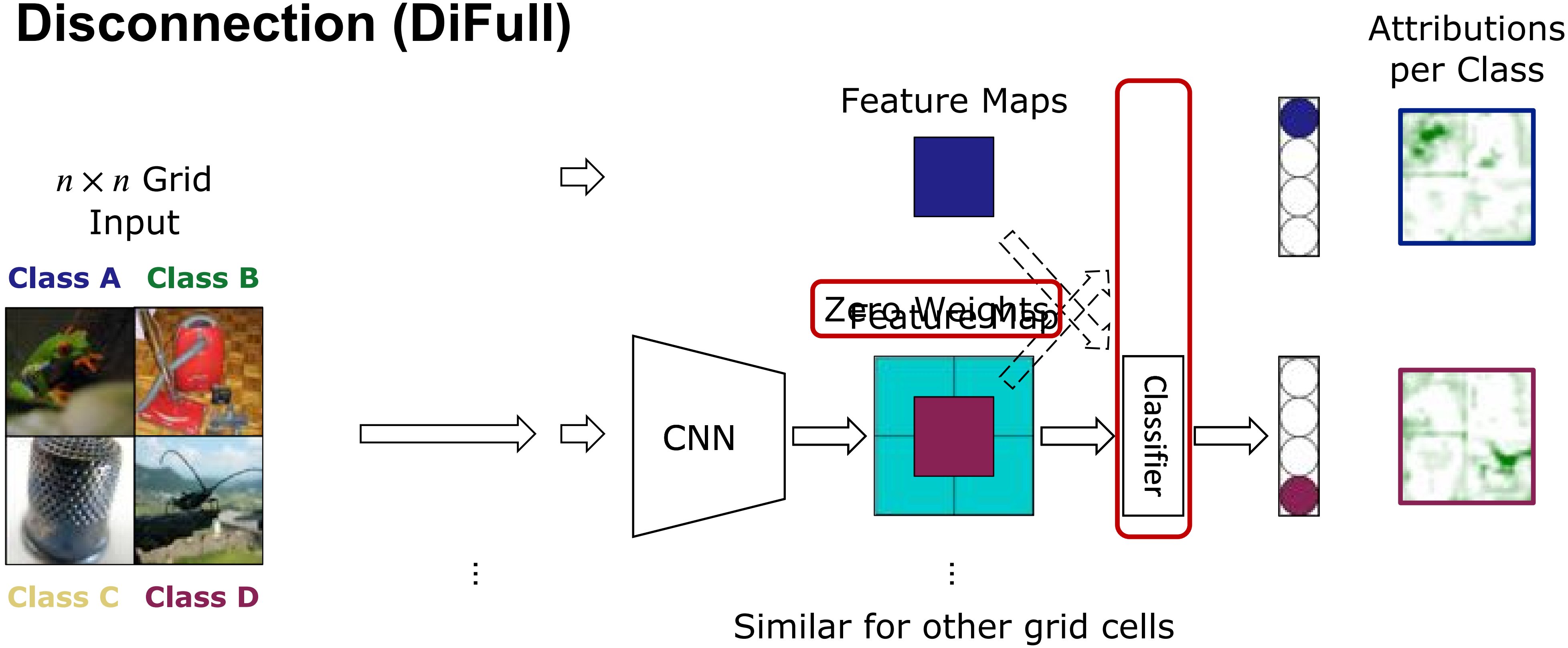
Semantically
Similar



Possibly
influential for
classifying
Class A

Need to disentangle model contribution from attribution method

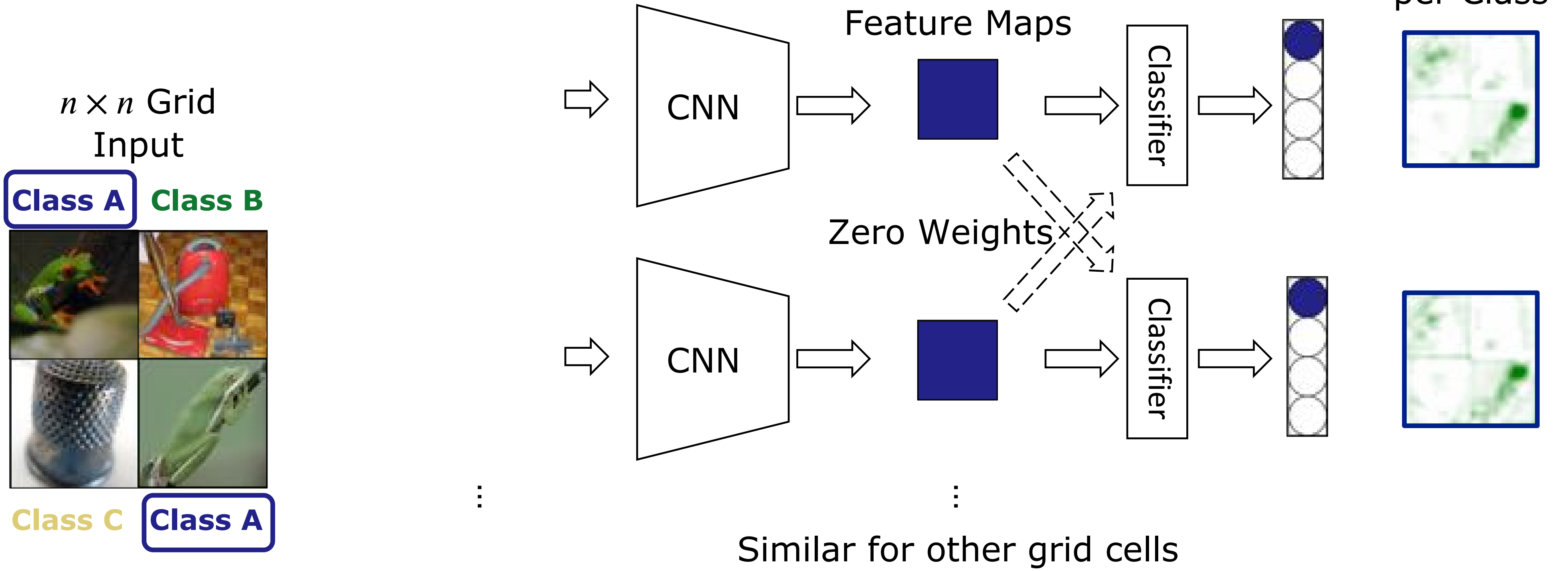
Full Disconnection (DiFull)



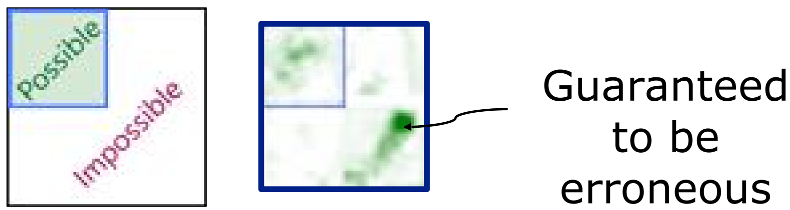
Guarantees that each classification head only influenced by its own grid cell

(For the top-left grid cell)

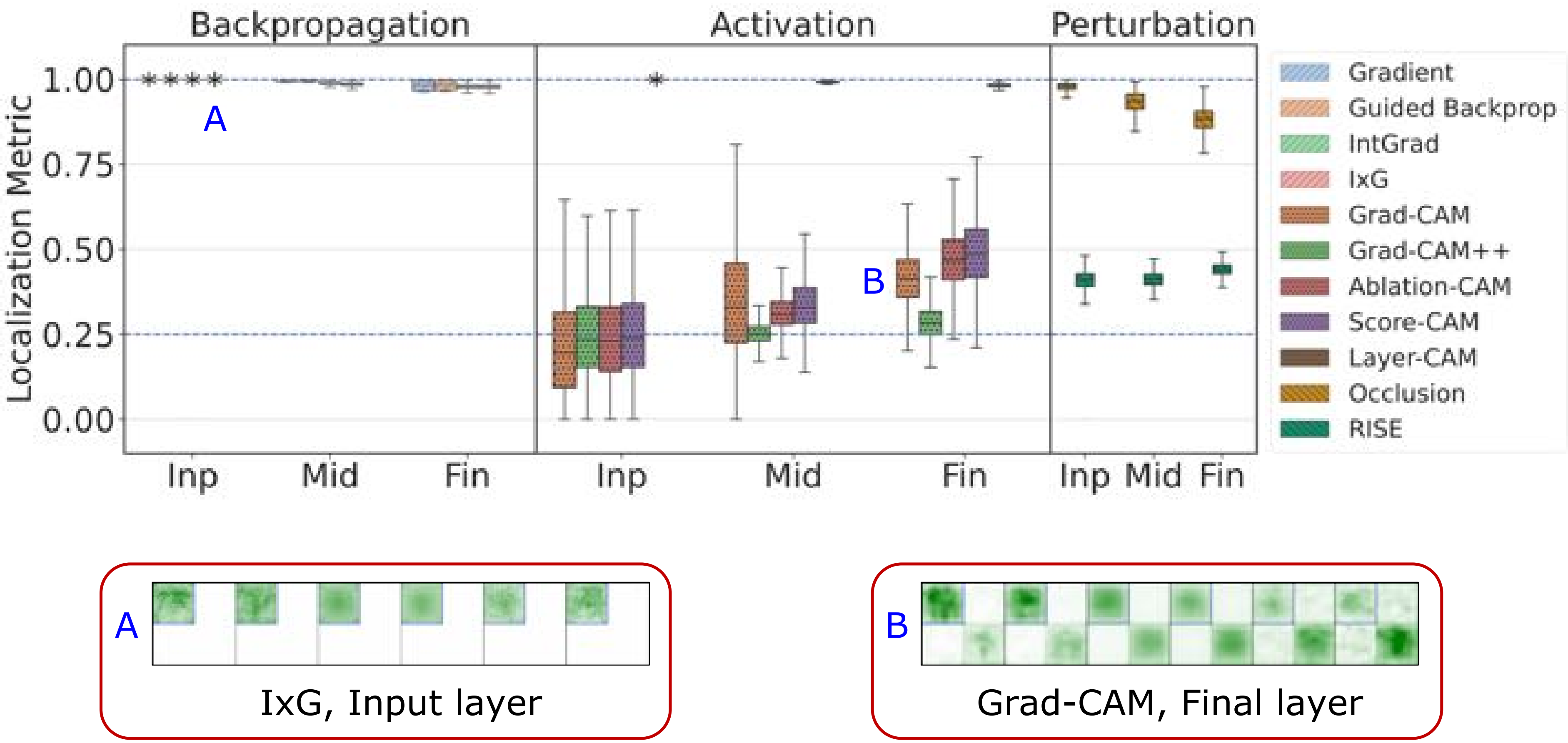
Full Disconnection (DiFull)



Stronger evaluation: Repeat class



Results: DiFull



Interim Summary — Post-Hoc Attribution Methods

- Difficult to evaluate post-hoc attribution methods
  - unknown ground-truth of model-contribution
  - difficult to disentangle of model contribution & attribution method

- Attribution at the last layers relatively easy
  - reason, why Grad-CAM is used widely
  - but only the very last layer(s) explained
  - can be very misleading - see our DiFull-setting

Overview

- **Interpretability** for Deep Learning in Computer Vision
  - Towards Better Understanding of Attribution Methods — CVPR’22, arXiv’23 [2303.11884]
  - **Inherently Interpretable** CNN Networks — CVPR’21, CVPR’22
  - **Inherently Interpretable** Transformer Networks — arXiv’23 [2301.08571]
  - Using **Explanations** to **Guide Inherently Interpretable** Models — ICCV’23

max planck institut
informatik

SIC Saarland Informatics Campus

CoDA-Nets: Convolutional Alignment Networks for Interpretable Classification

@ CVPR 2021

B-cos Networks: Alignment is All We Need for Interpretability

@ CVPR 2022

Moritz Boehle
MPI Informatics

Mario Fritz
CISPA Helmholtz

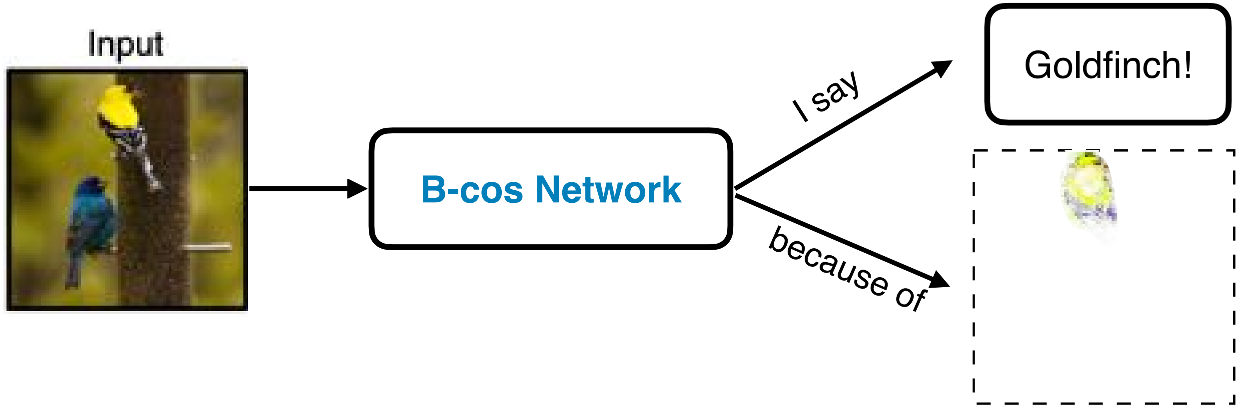
Bernt Schiele
MPI Informatics

Motivation: we aim for Inherent Interpretability

Want: Explanation

Requirements: Faithful ✓ | Interpretable ✓

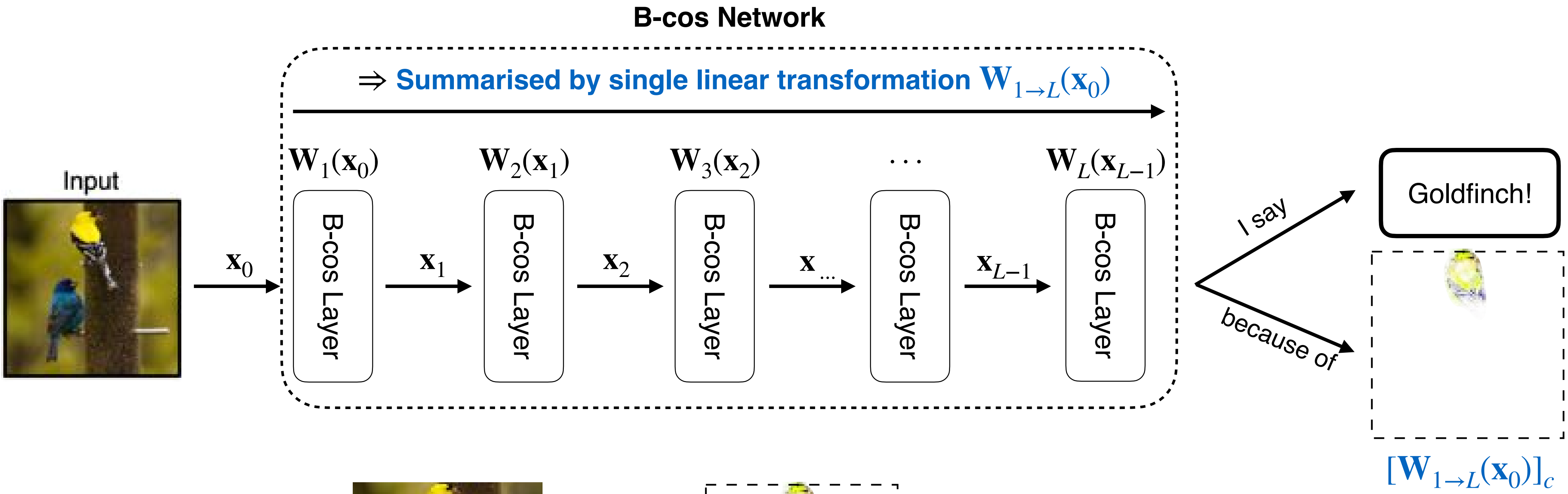
Dynamic linearity | **Alignment pressure**

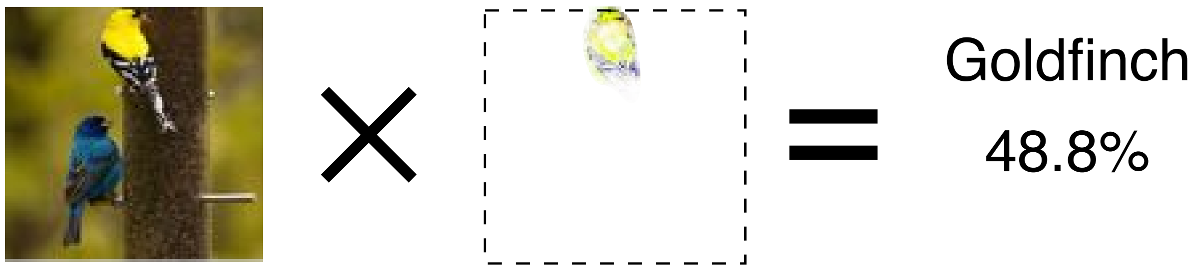


Model-inherent linear map

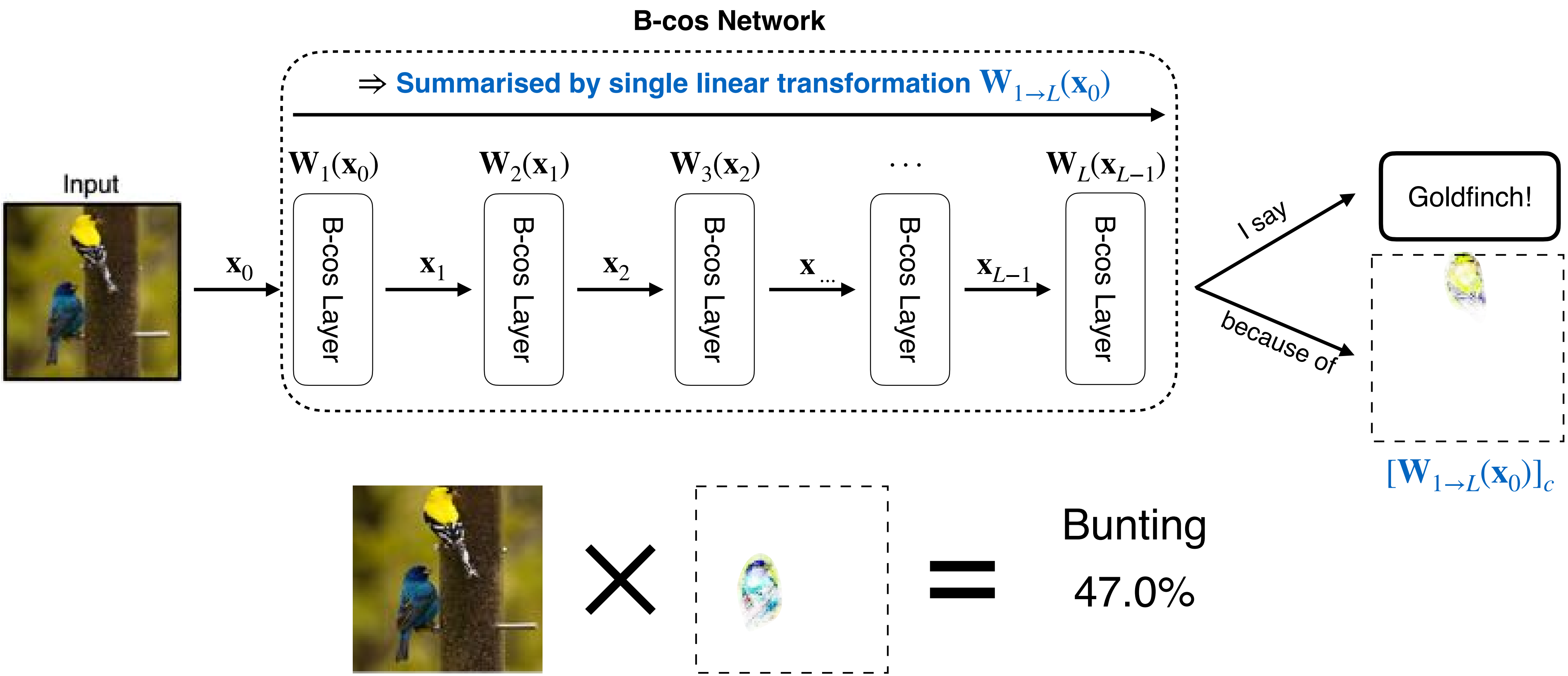
References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

B-cos Networks: Dynamic Linearity

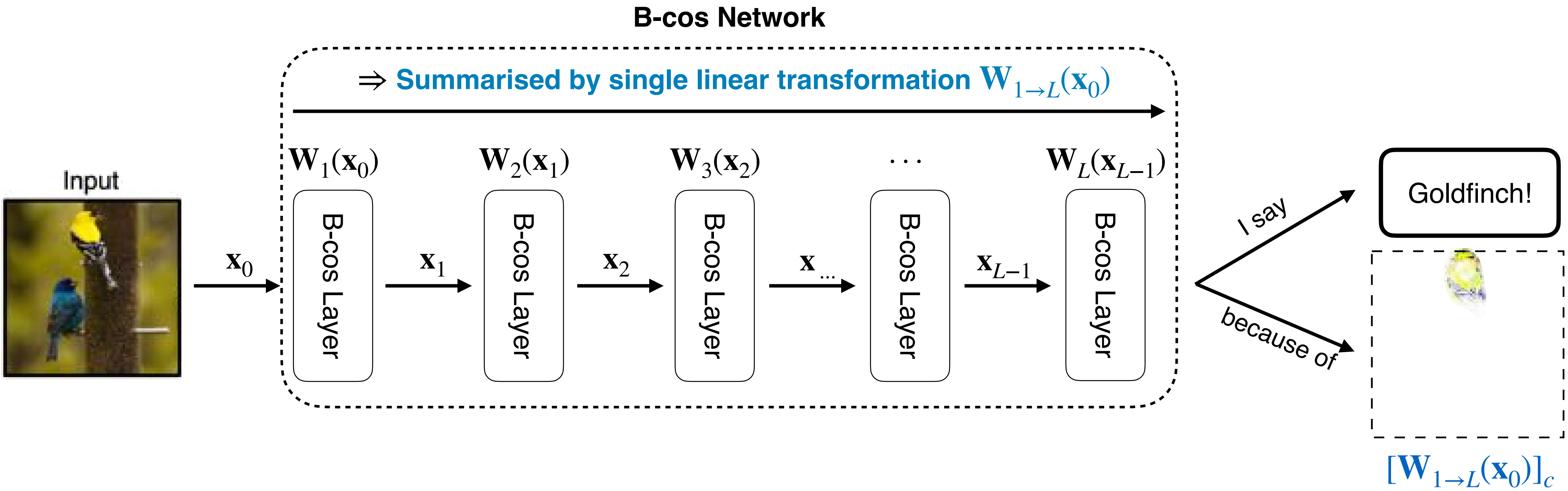




B-cos Networks: Dynamic Linearity



B-cos Networks: Dynamic Linearity



Dynamic linearity allows us to faithfully summarise the model.

Alignment pressure

B-cos transformation vs. linear transformation

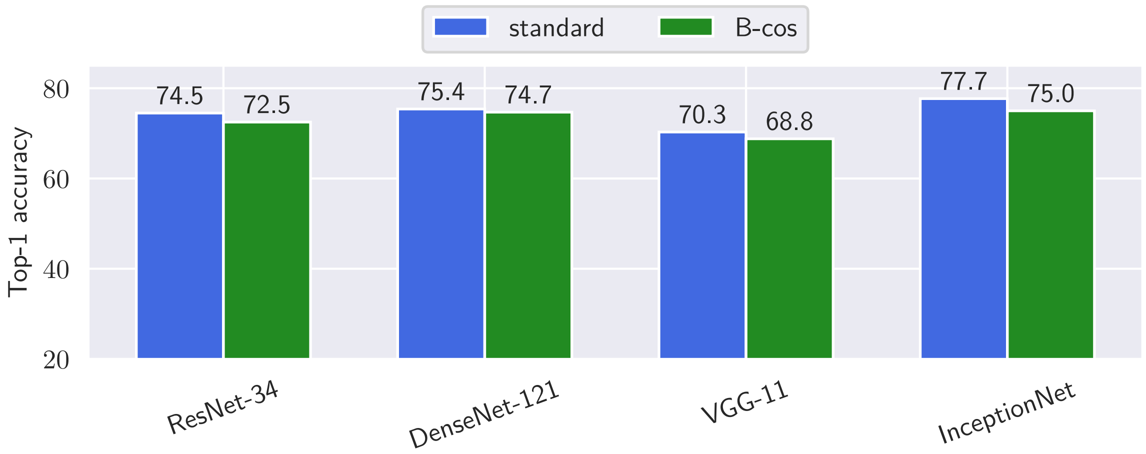
Linear transformation $f(\mathbf{x};\ \mathbf{w}) = \mathbf{w}^T\ \mathbf{x} = ||\mathbf{w}||\ ||\mathbf{x}||\ \cos(\mathbf{x}, \mathbf{w})$

New transformation $\text{B-cos}(\mathbf{x};\ \mathbf{w}) = \underbrace{||\ \widehat{\mathbf{w}}\ ||}_{=1}\ ||\mathbf{x}||\ |\cos(\mathbf{x}, \mathbf{w})|^{B} \times \text{sgn}\left(\cos(\mathbf{x}, \mathbf{w})\right)$

ImageNet results

Compatible with standard architectures

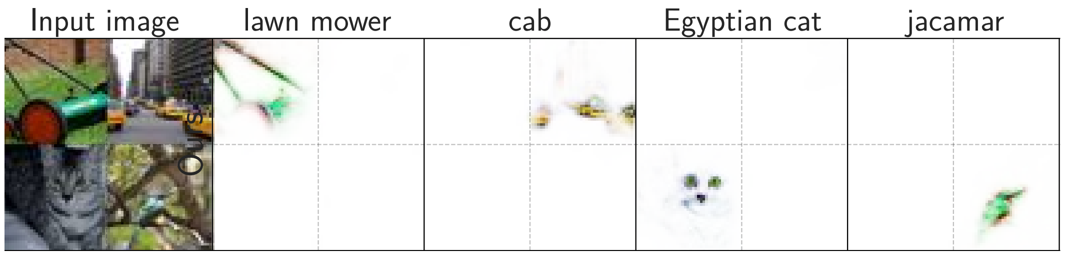
B-cos networks achieve competitive accuracies



Measuring Interpretability via Grid Pointing Game

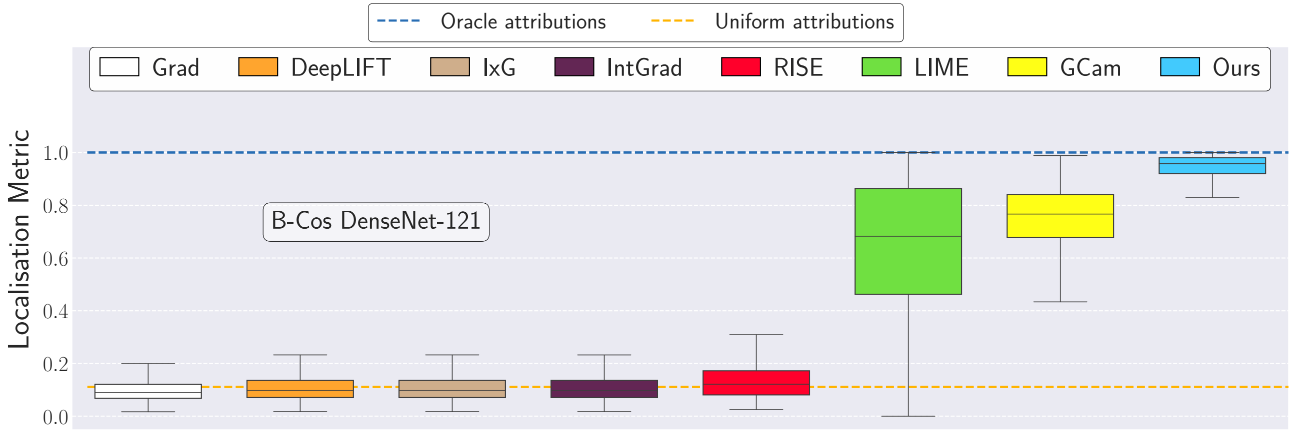
- To measure interpretability, we employ the *grid pointing game*

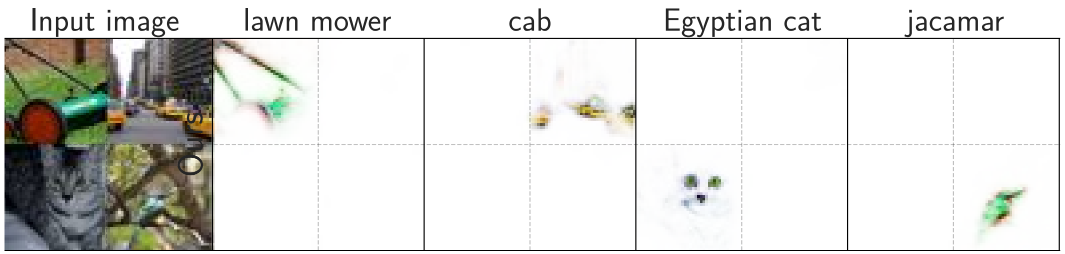
- In particular:
  - evaluate models on synthetic image grid
  - measure how well an explanation *localises* the correct image grid
  (score $s = \frac{A_i^+}{\sum_j A_j^+}$ with A_i^+ the positive attribution to subimage i)



ImageNet results

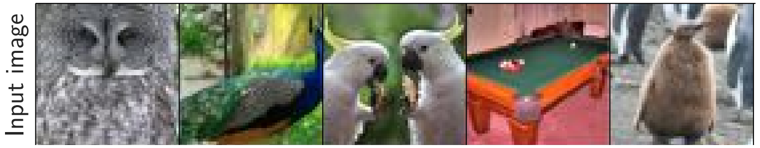
High interpretability



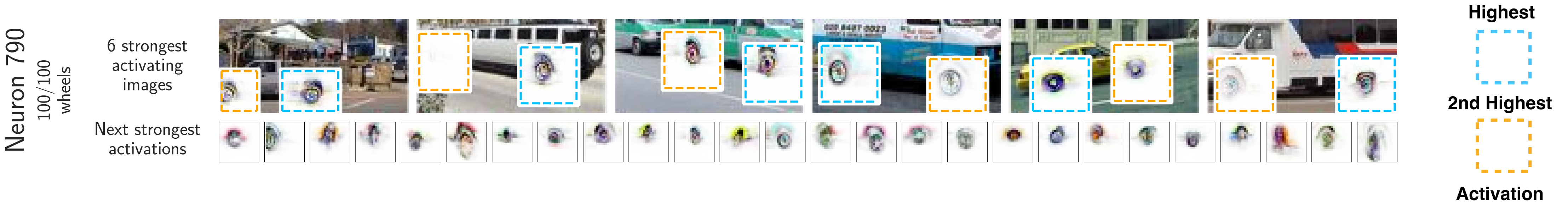


Gradient (Baehrens (2010)), DeepLIFT (Shrikumar (2017)), Input x Gradient (cf. Adebayo (2018)), IntGrad (Sundararajan (2017)), RISE (Petsiuk, 2018), LIME (Ribeiro, 2016), GradCam (Ramprasaath et al. (2017))

Visualisations of $\mathbf{W}_{1\rightarrow L}(\mathbf{x})$



Visualisations: intermediate neurons



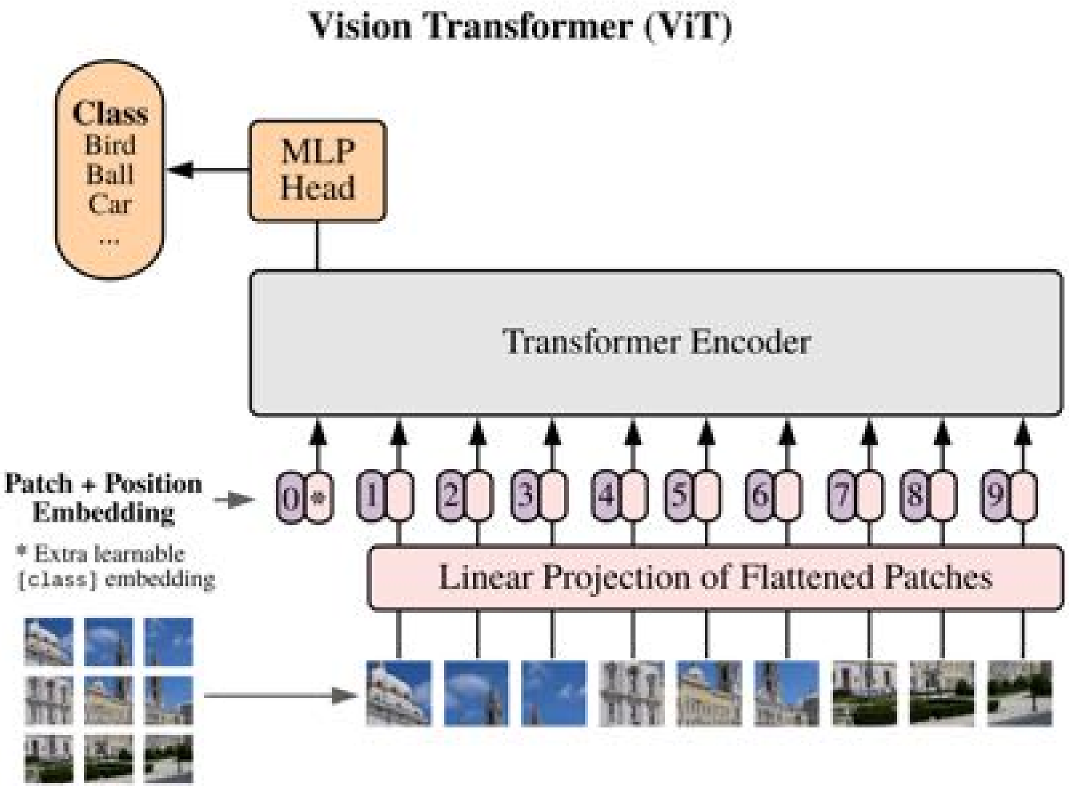
Interim Summary

- Deep Neural Network explanations need to be **faithful** & **interpretable**
  - for faithfulness: B-cos is designed to be **dynamic linear**
  - for interpretability: B-cos induces **alignment pressure**
- The resulting networks are **competitive classifiers**...
- ... and **provide interpretable explanations** for their decisions

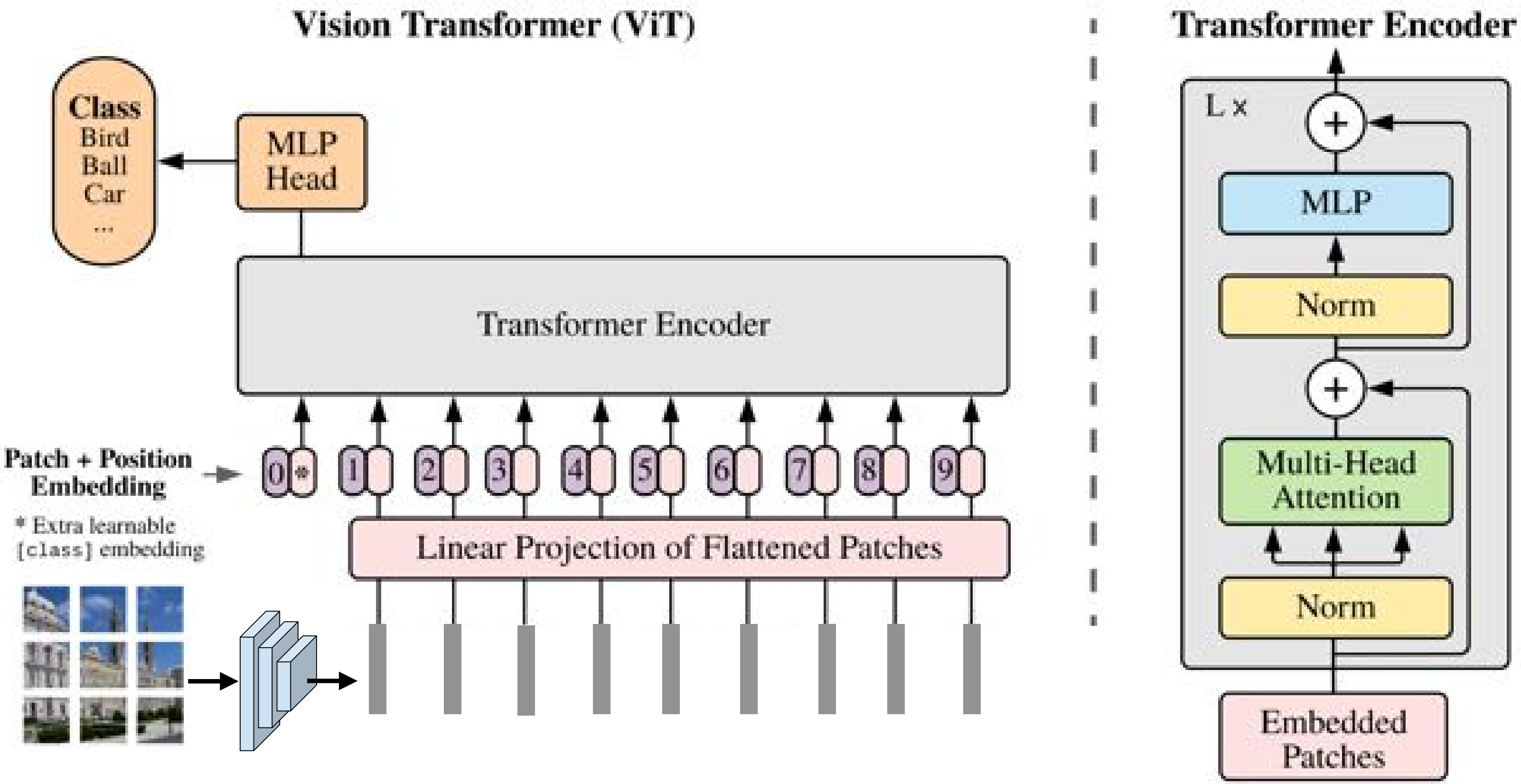
Overview

- **Interpretability** for Deep Learning in Computer Vision
  - Towards Better **Understanding** of **Attribution** Methods — CVPR’22, arXiv’23 [2303.11884]
  - **Inherently Interpretable** CNN Networks — CVPR’21, CVPR’22
  - **Inherently Interpretable** Transformer Networks — arXiv’23 [2301.08571]
  - Using **Explanations** to **Guide Inherently Interpretable** Models — ICCV’23

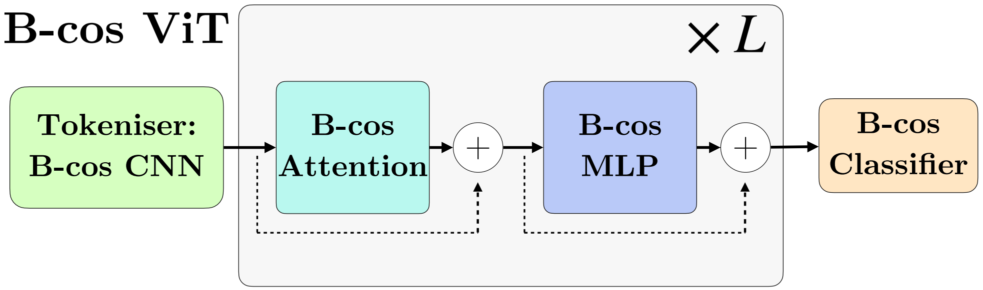
Attention is not All You Need (for XAI)



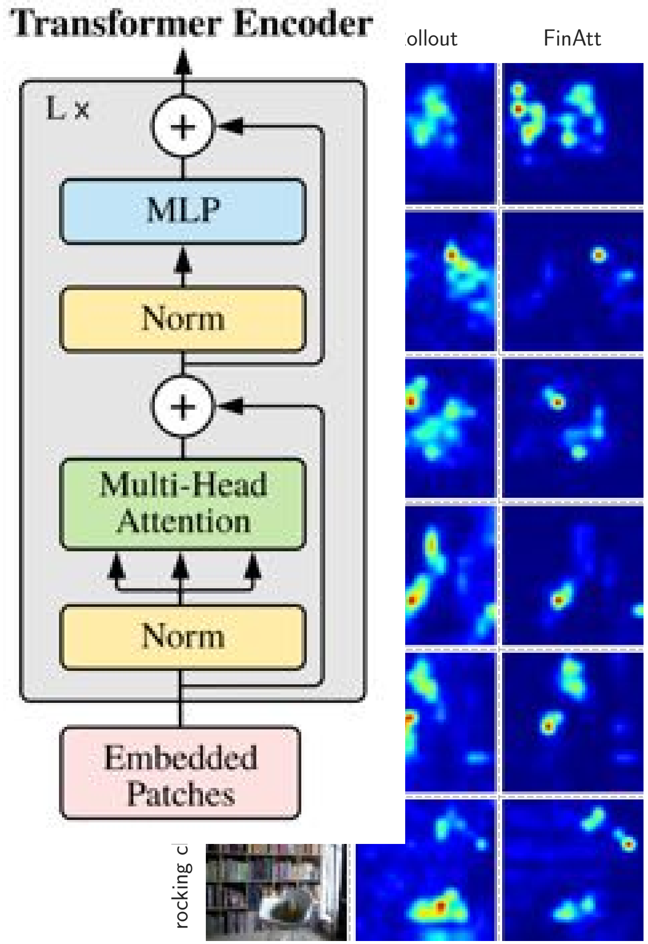
Attention is not All You Need (for XAI)



Attention is not All You Need (for XAI)

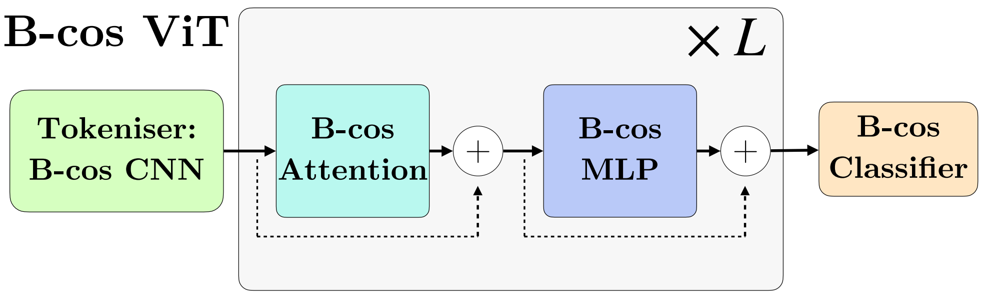


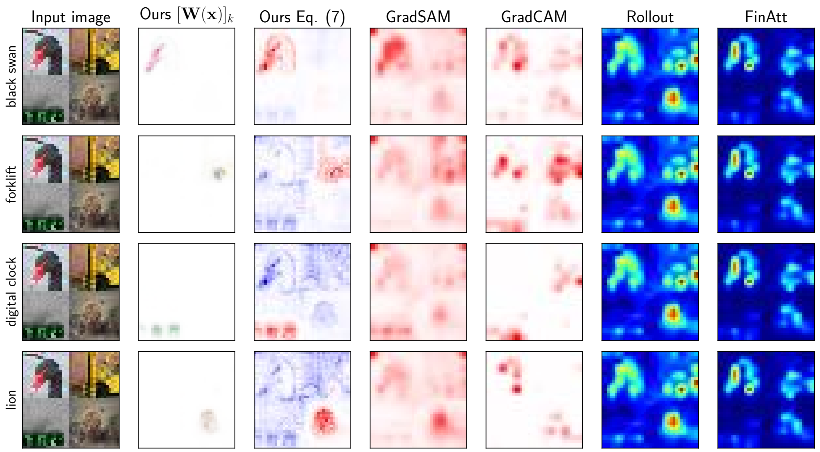
- High resoluti
- Holistic expla

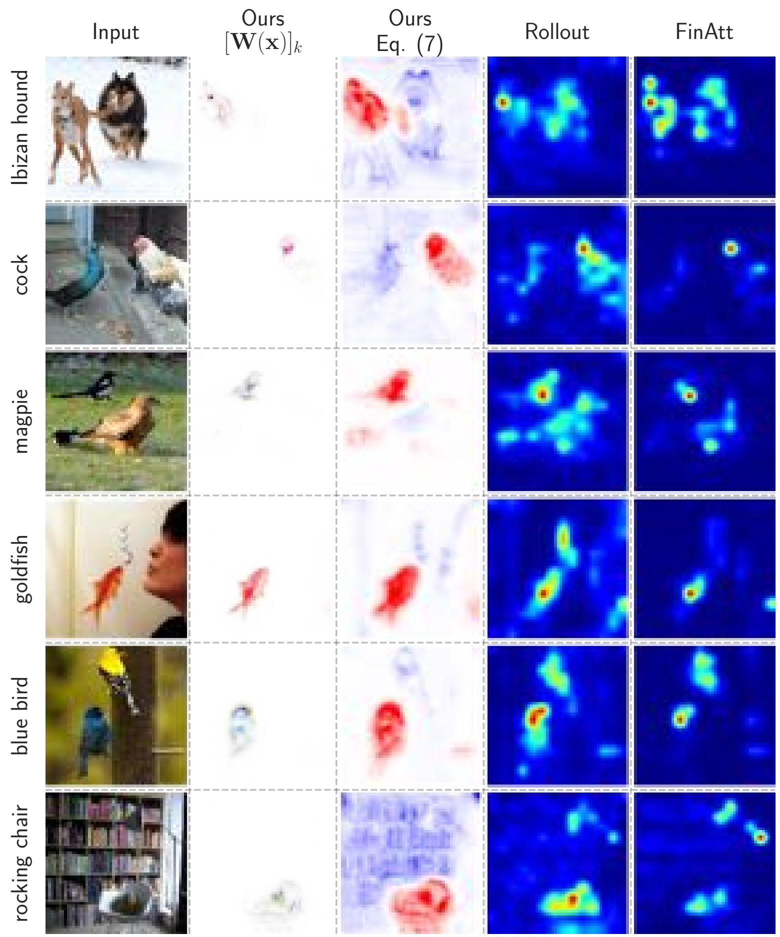


Attention Rollout: Abnar & Zuidema, ACL 2020

Attention is not All You Need (for XAI)



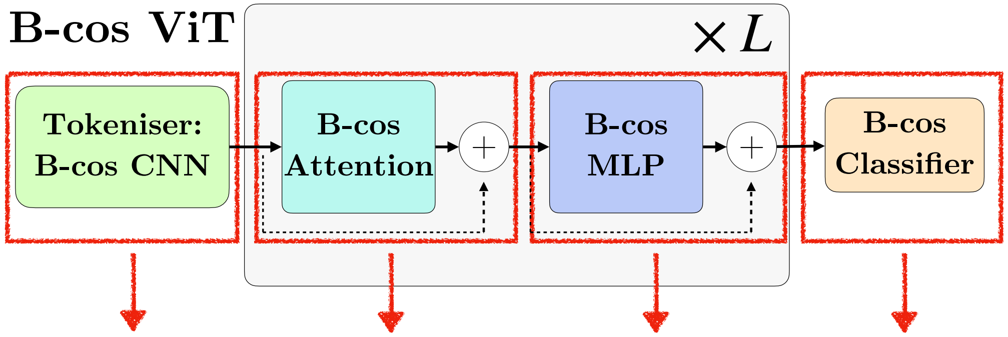




Attention Rollout: Abnar & Zuidema, ACL 2020; GradSAM: Barkan et al., CIKM 2021, GradCAM: Selvaraju et al., ICCV 2017

Attention is not All You Need (for XAI)

- Tokeniser + MLP + Classifier
  - interpret as CNNs,
    convert to B-cos CNNs
- Self-Attention (SA) is dynamic linear
  - $$SA(\mathbf{X}) = \underbrace{\mathbf{A}(\mathbf{X})\,\mathbf{V}}_{\mathbf{W}(\mathbf{X})}\ \mathbf{X} = \mathbf{W}(\mathbf{X})\,\mathbf{X}$$
- For this talk:
  - for Tokenisation, use L layers of
    pretrained+frozen B-cos DenseNet-121



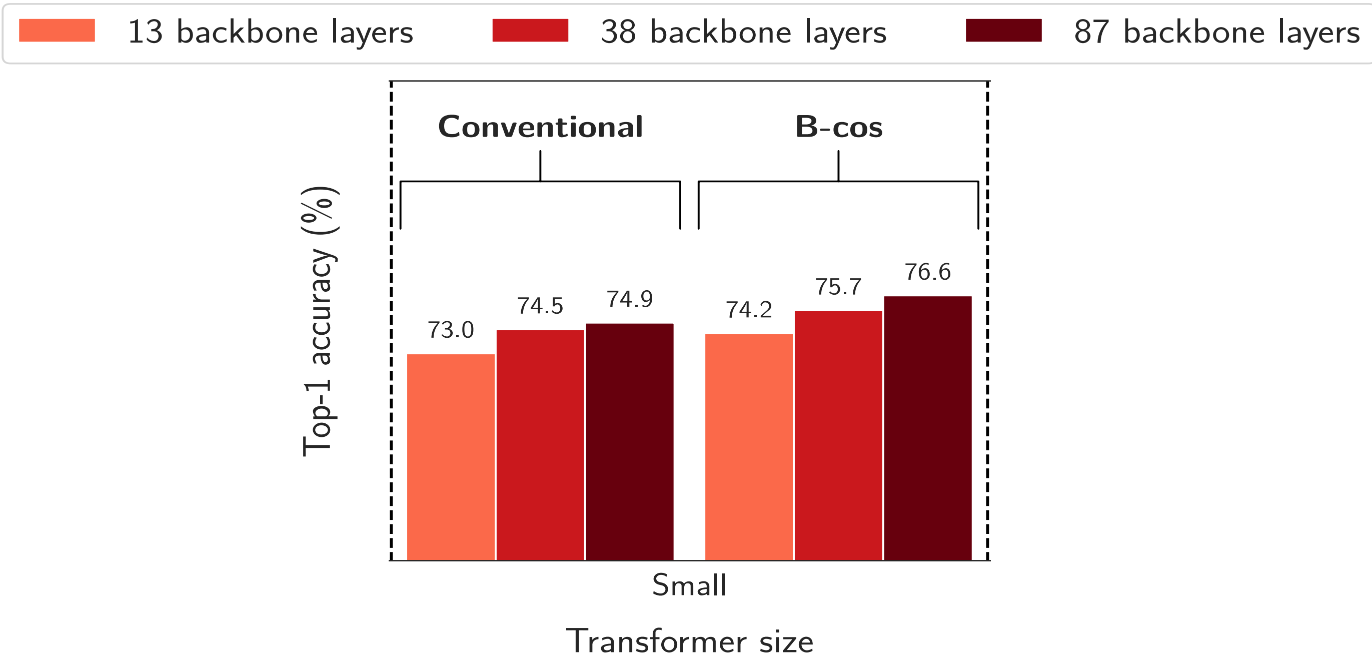
$$\mathbf{W}^{\text{Tokens}}(\mathbf{x}) \quad \mathbf{W}_l^{\text{Att}}(\mathbf{x}) \quad \mathbf{W}_l^{\text{MLP}}(\mathbf{x}) \quad \mathbf{W}^{\text{Class}}(\mathbf{x})$$

Dynamic Linear Transformation

$$\mathbf{W}(\mathbf{x}) = \mathbf{W}^{\text{Class}}(\mathbf{x}) \prod_{l=1}^{L}\left(\mathbf{W}_l^{\text{MLP}}(\mathbf{x})\,\mathbf{W}_l^{\text{Att}}(\mathbf{x})\right)\ \mathbf{W}^{\text{Tokens}}(\mathbf{x})$$

Attention is not All You Need (for XAI)

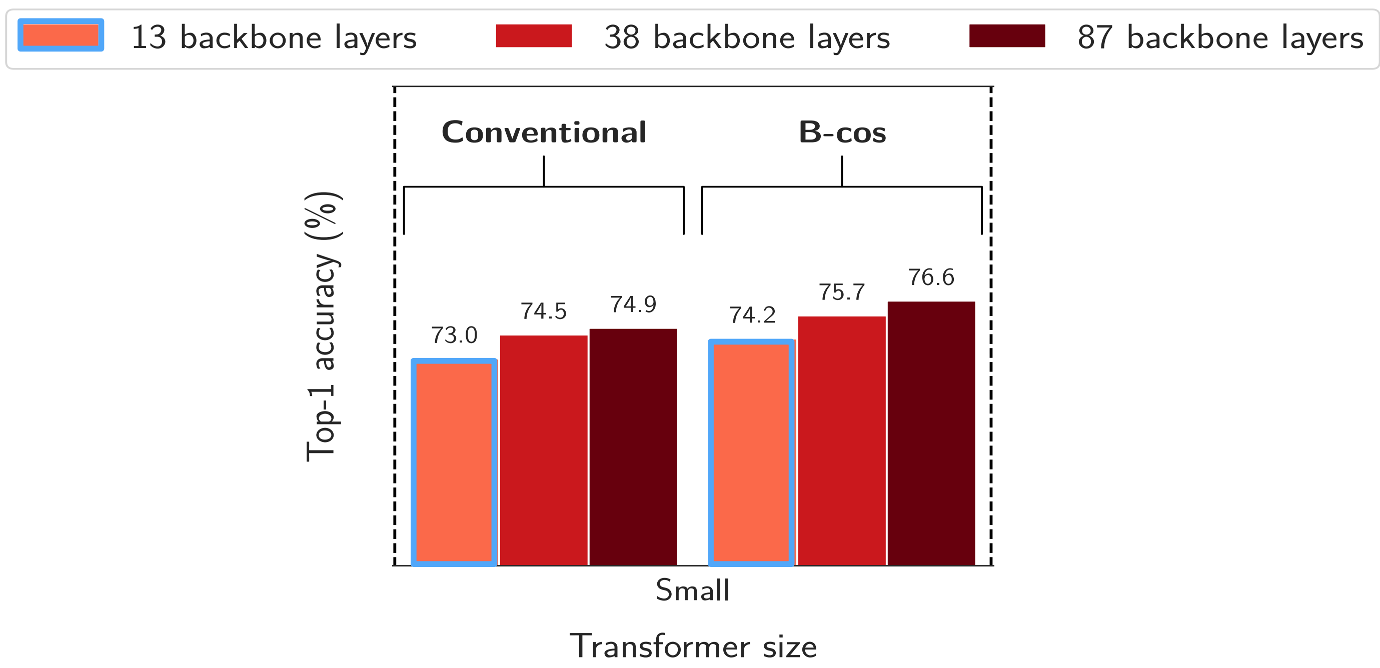
Results — classification accuracy



Attention is not All You Need (for XAI)

- Results

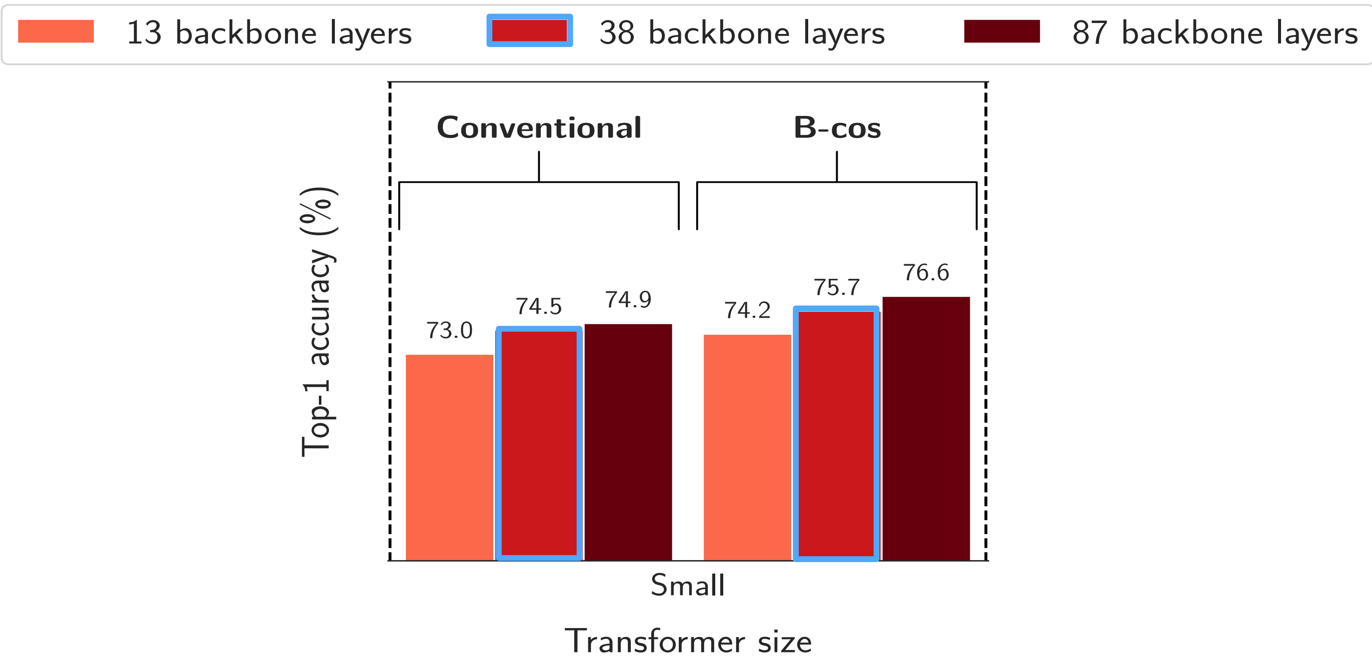
Results — classification accuracy



Attention is not All You Need (for XAI)

- Results

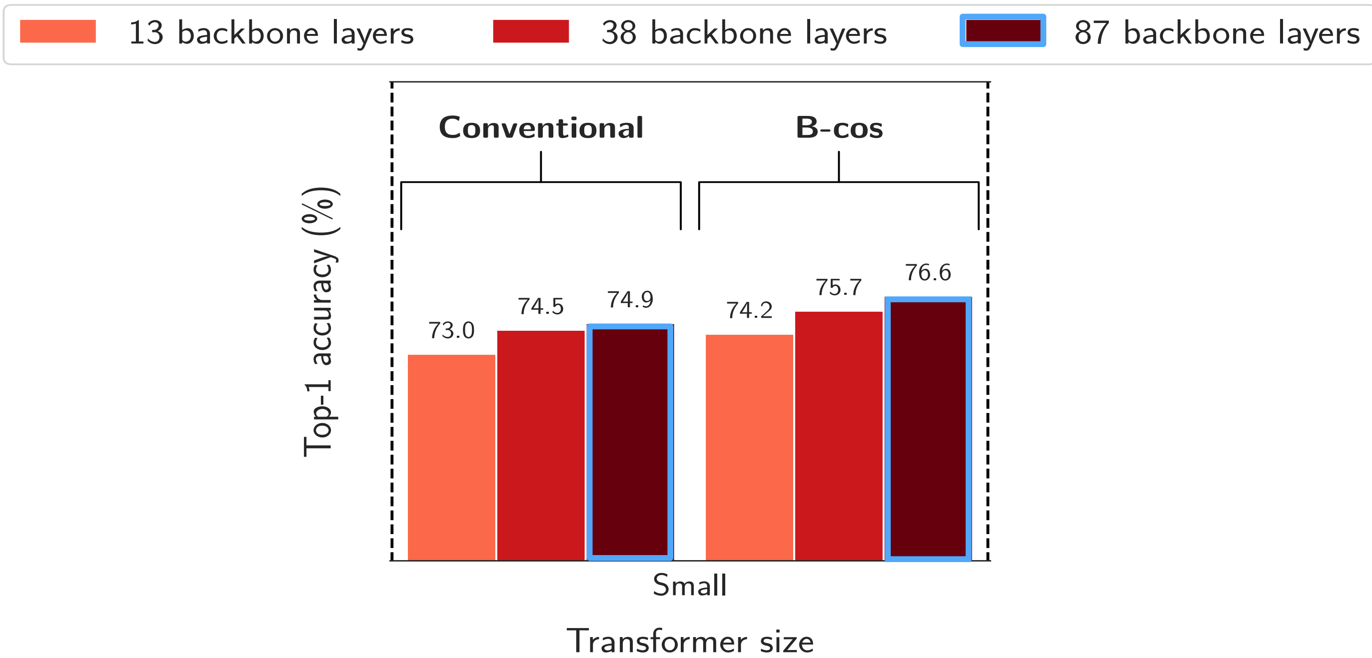
Results — classification accuracy



Attention is not All You Need (for XAI)

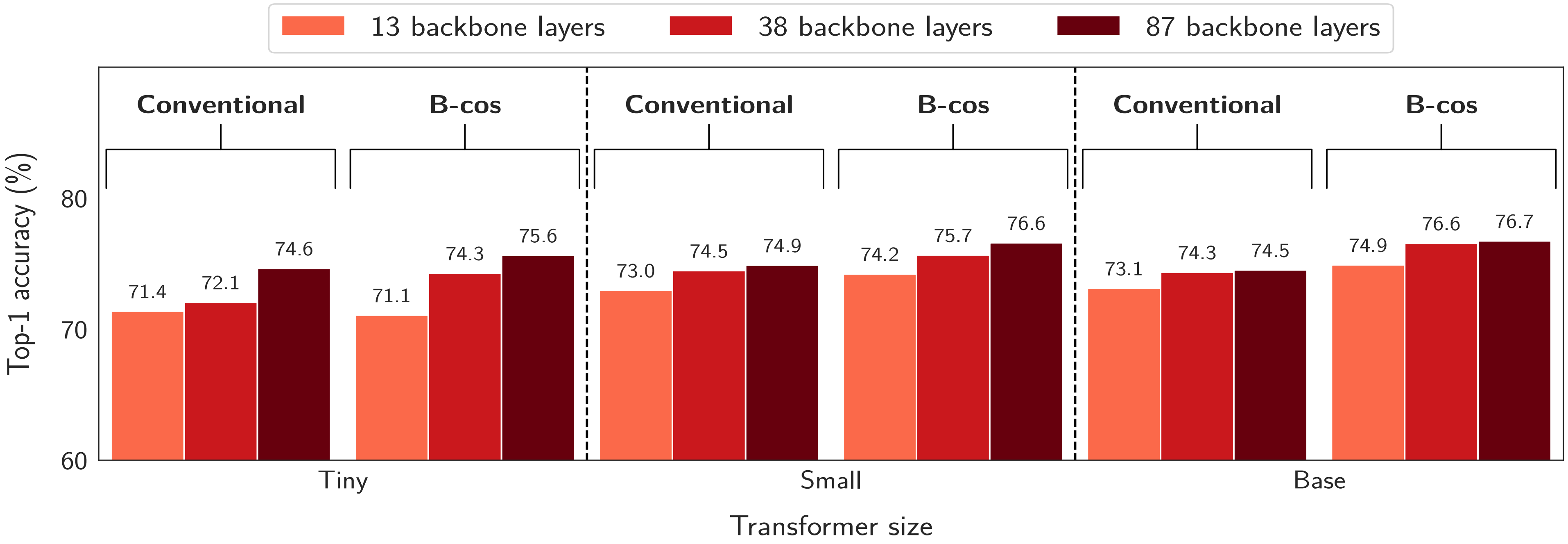
- Results

Results — classification accuracy

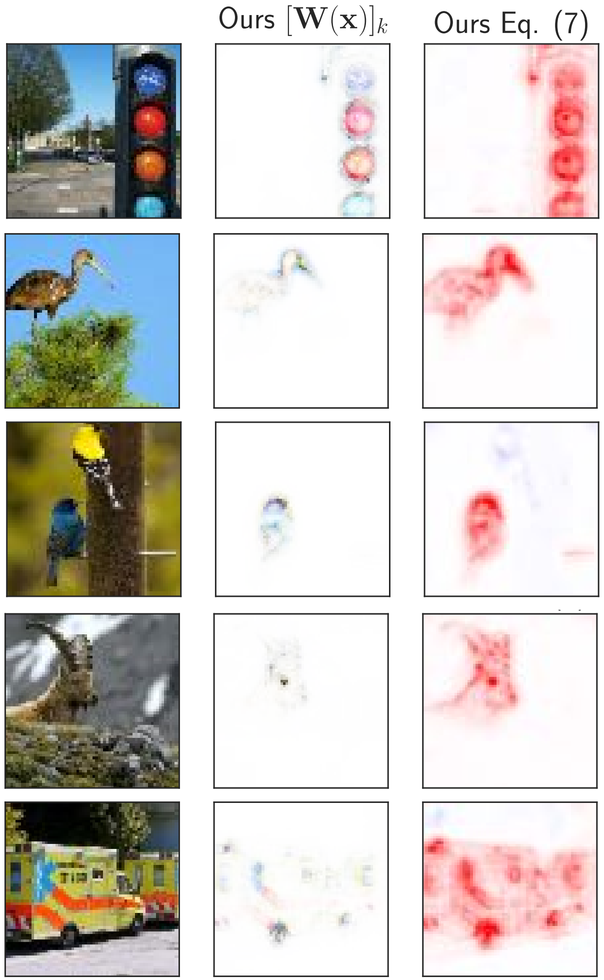


Attention is not All You Need (for XAI)

Results — classification accuracy

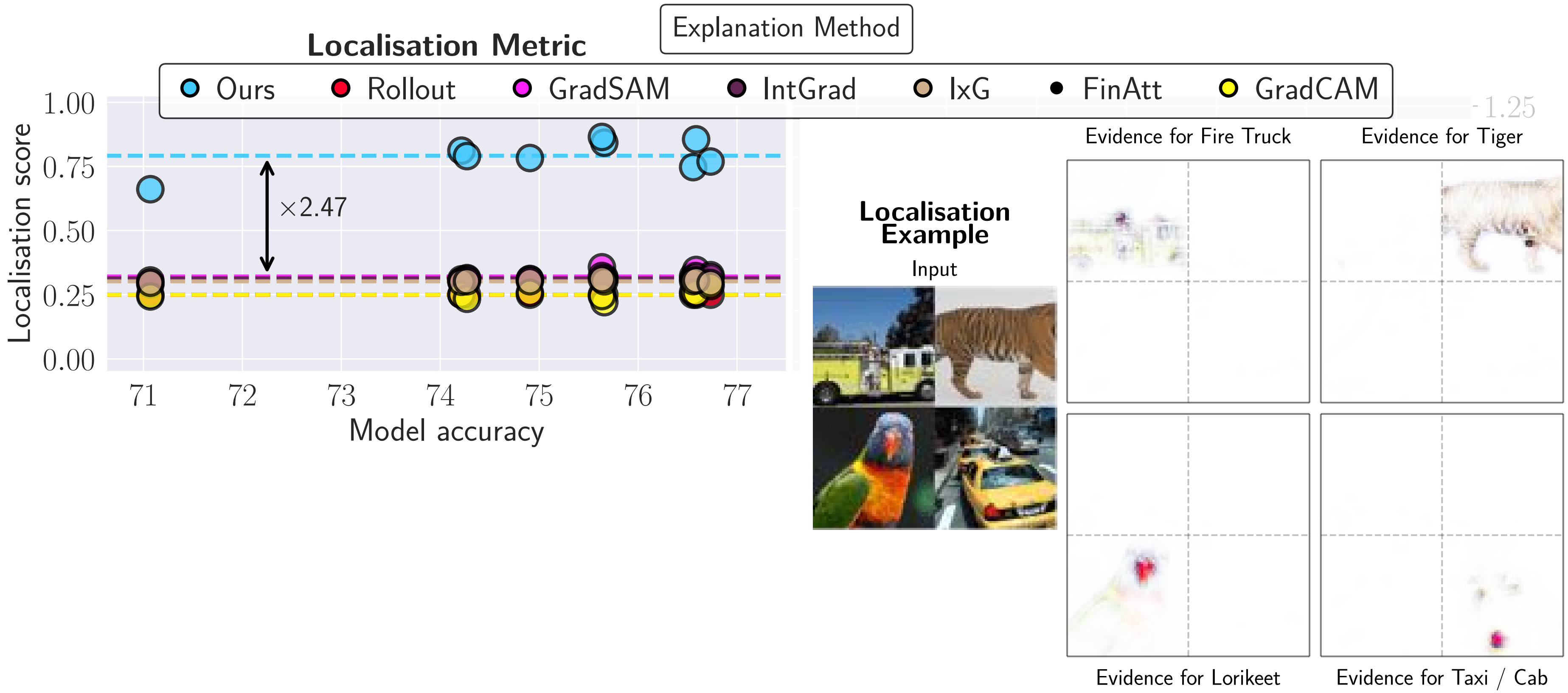


Qualitative Results — Interpretability



Attention is not All You Need (for XAI)

Results — interpretability metrics



Attention is not All You Need (for XAI) - Summary

- B-cos framework generally compatible with ViTs
  - Attention already dynamic linear SA$(\mathbf{X}) = \mathbf{W}(\mathbf{X})\mathbf{X}$
  - remaining modules $\rightarrow$ B-cos CNNs
- B-cos ViTs can be highly performant
  - similar results as with standard ViTs in comparable setting
- B-cos ViTs highly interpretable
  - similar interpretability as B-cos CNNs

Overview

- **Interpretability** for Deep Learning in Computer Vision
  - Towards Better **Understanding** of **Attribution** Methods — CVPR'22, arXiv'23 [2303.11884]
  - **Inherently Interpretable** CNN Networks — CVPR'21, CVPR'22
  - **Inherently Interpretable** Transformer Networks — arXiv'23 [2301.08571]
  - Using **Explanations** to **Guide Inherently Interpretable** Models — ICCV'23

Using Explanations to Guide Models

@ ICCV'23




Sukrut Rao*
MPI Informatics

Moritz Boehle*
MPI Informatics

Amin Parchami-Araghi
MPI Informatics

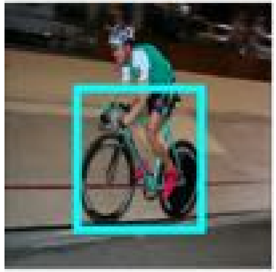
Bernt Schiele
MPI Informatics

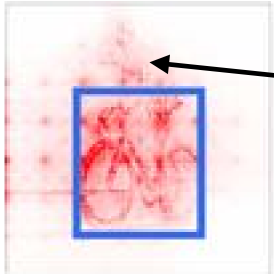
Motivation

- Deep networks may rely on spurious features

Image | Attribution | | Problem

Bicycle

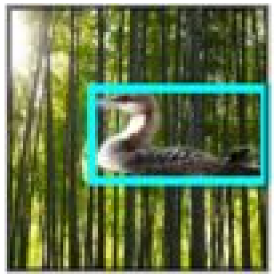


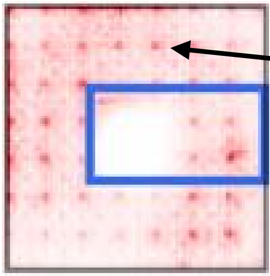


Uses commonly co-occurring feature

Poor Generalization

Waterbird



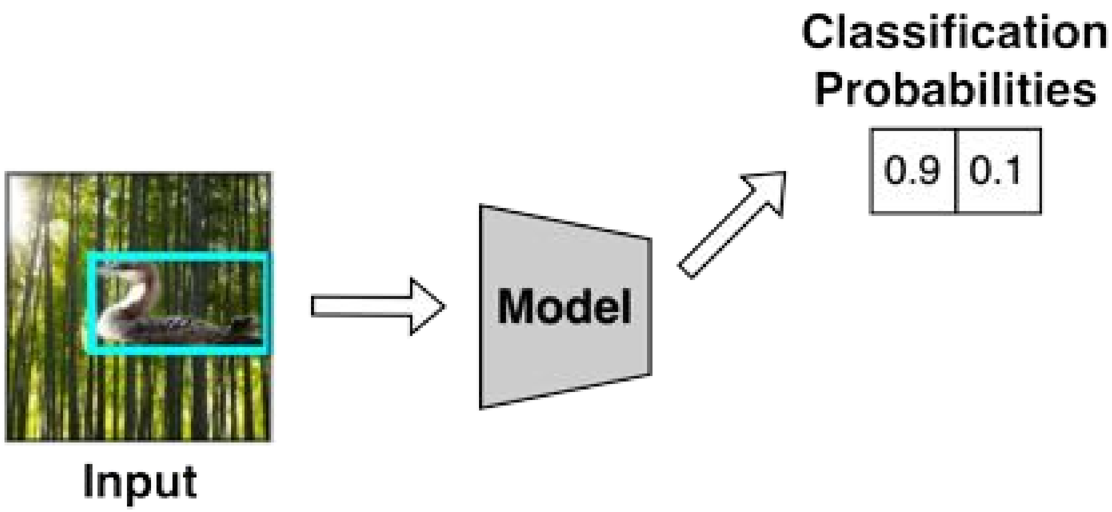


Uses spurious background features

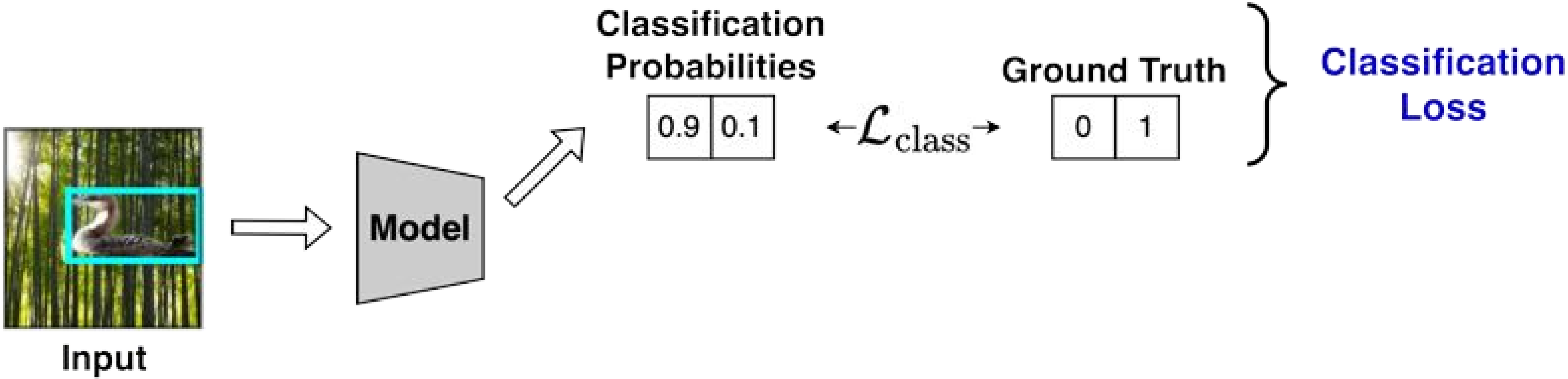
Misclassification

- **Idea: Guide models to be "right for the right reasons"**

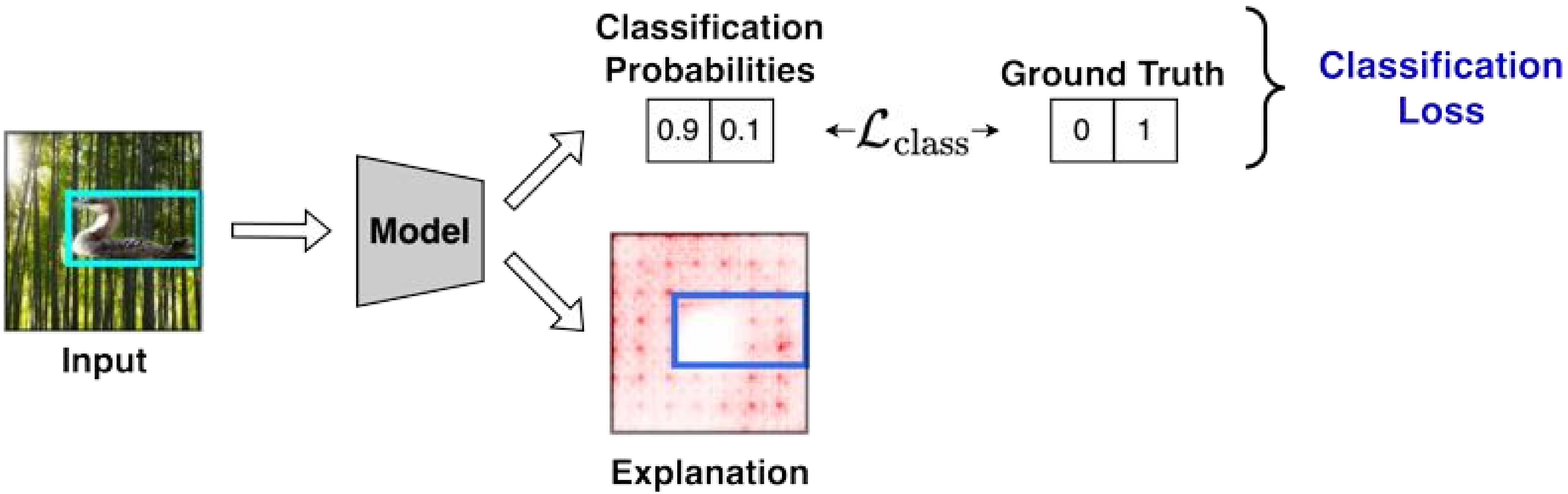
Guiding Models: Pipeline



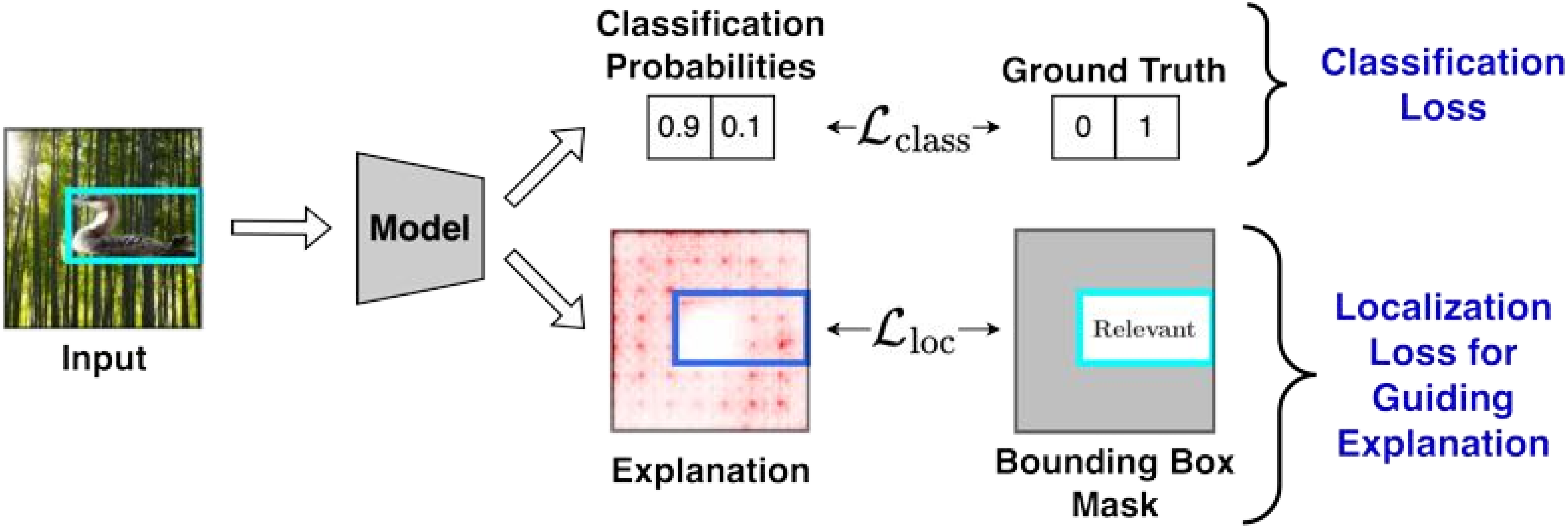
Guiding Models: Pipeline



Guiding Models: Pipeline

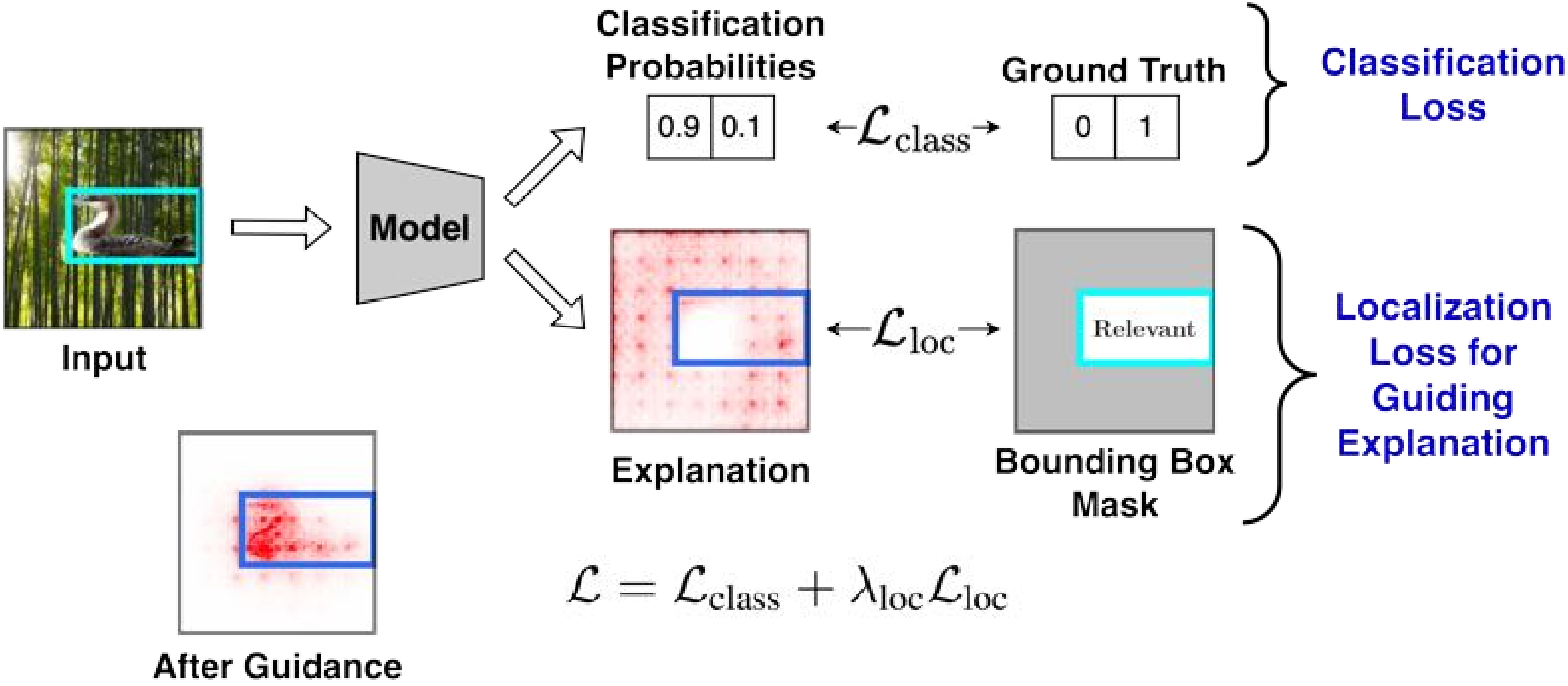


Guiding Models: Pipeline



$$\mathcal{L} = \mathcal{L}_{\text{class}} + \lambda_{\text{loc}} \mathcal{L}_{\text{loc}}$$

Guiding Models: Pipeline



$$\mathcal{L} = \mathcal{L}_{\text{class}} + \lambda_{\text{loc}}\mathcal{L}_{\text{loc}}$$

Related Work

- **Explicit Guidance:** Specify where the model should look
- **Forms of Guidance:**
  - Language Models (e.g. GALS [Petryk et al., 2022]):

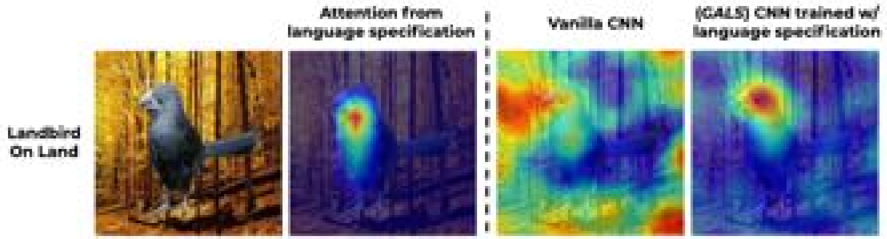


Image Source: Petryk et al. On Guiding Visual Attention with Language Specification. CVPR 2022

  - Image Annotation Masks (e.g. RES [Gao et al., 2022]):

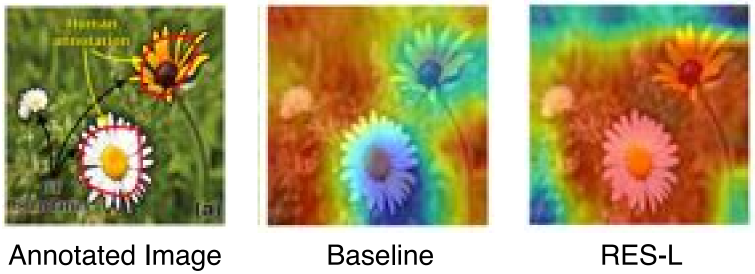


Image Source: Gao et al. RES: A Robust Framework for Guiding Visual Explanation. KDD 2022

Our Focus

With *coarse* annotation masks, i.e. bounding boxes

Related Work: Guidance with Annotations

- **Datasets** are often:
  - **Small:** a few hundred or thousand images
  - **Simple:** binary classification
  - **Synthetic:** constructed, often not using natural images
- **Attribution methods**:
  - Fixed, usually GradCAM
  - Coarse grained, explain only the final layer
- **Localization losses**:
  - Often (e.g. with L_1 loss) enforce uniformity in mask

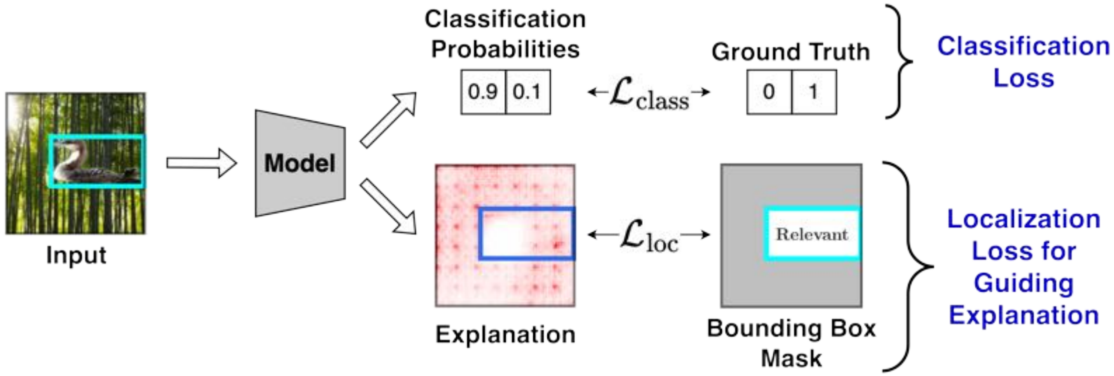
Ours

Large scale, multi-label classification datasets (PASCAL VOC, MS COCO)

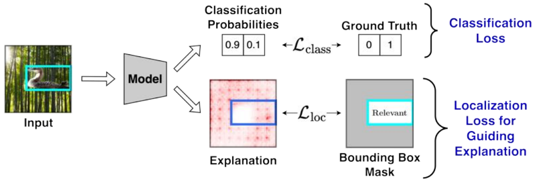
Diverse set of attribution methods at multiple depths

Novel Energy loss, comparison against multiple loss functions

Localization Losses from Prior Work



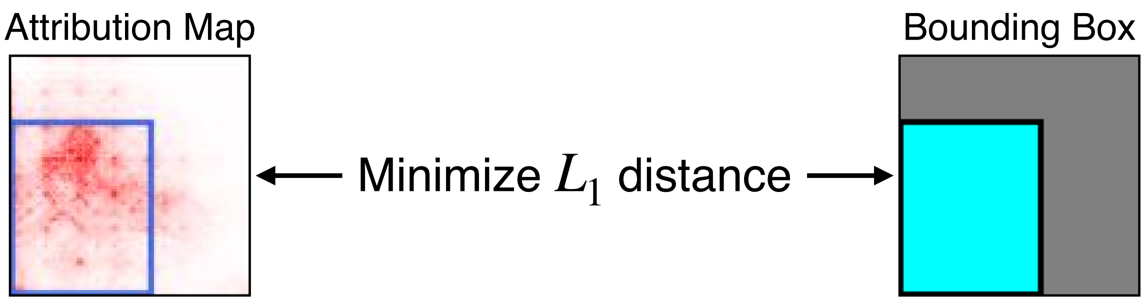
- Localization Losses from Prior Work:
  - L_1 Loss
  - Per-pixel Cross Entropy (PPCE) Loss
  - RRR* Loss (extended from RRR Loss)



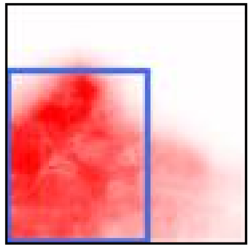
Localization Losses from Prior Work

- L_1 Loss:
  - Minimize L_1 distance between normalised attributions and annotation
  - Guides model to attribute uniformly to existing highest attribution value

Image



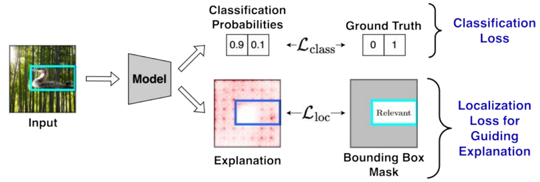
After Guidance



Attributions distributed uniformly in box

$$\mathcal{L}_{\mathrm{loc},k} = \frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}\|M_{k,hw} - \hat{A}^{+}_{k,hw}\|_1$$

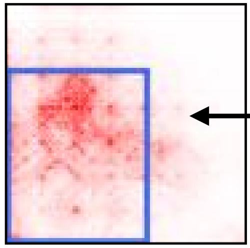
Localization Losses from Prior Work



- Per-pixel cross entropy (PPCE) Loss:
  - Use cross-entropy loss at every pixel inside bounding box
  - No explicit constraint on attributions outside the box

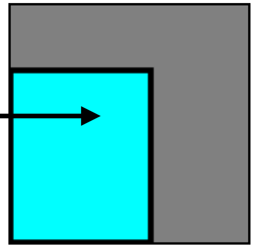
Image

Attribution Map

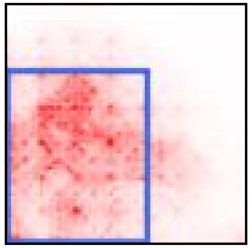


Cross entropy at each pixel *inside* box

Bounding Box



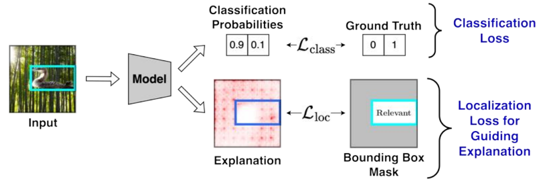
After Guidance



Not very effective

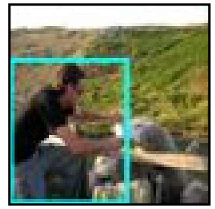
$$\mathcal{L}_{\text{loc},k} = -\frac{1}{\|M_k\|_1} \sum_{h=1}^{H} \sum_{w=1}^{W} M_{k,hw} \log(\hat{A}^{+}_{k,hw})$$

Localization Losses from Prior Work

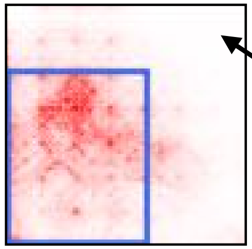


- RRR* Loss:
  - Minimizes square of attributions outside box

Image

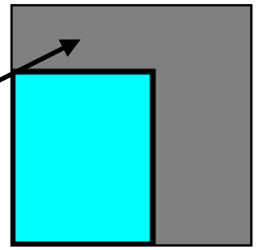


Attribution Map

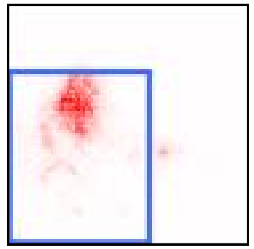


Suppress attributions *outside* box

Bounding Box



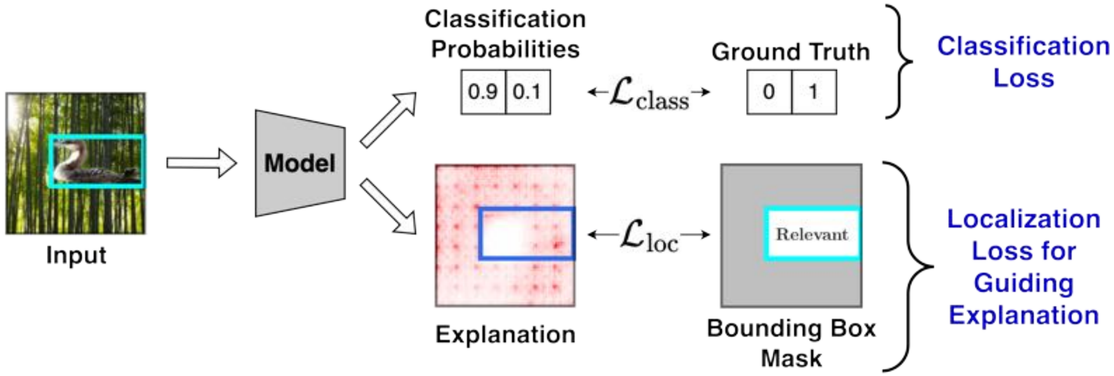
After Guidance



Sparse attributions, localizes well qualitatively, not as effective quantitatively

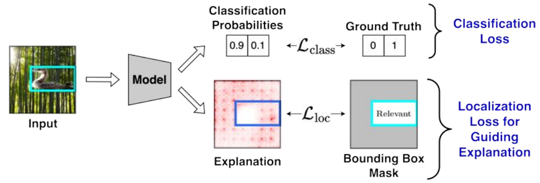
$$\mathcal{L}_{\text{loc},k} = \sum_{h=1}^{H} \sum_{w=1}^{W} (1 - M_{k,hw}) \hat{A}^2_{k,hw}$$

Localization Losses from Prior Work



- Localization Losses from Prior Work:
  - L_1 Loss
  - Per-pixel Cross Entropy (PPCE) Loss
  - RRR* Loss (extended from RRR Loss)
- **Ours:** Energy Loss

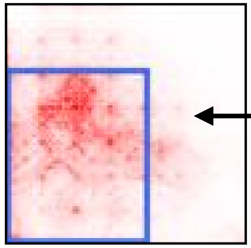
Energy Loss



- Energy Loss
  - Maximize fraction of attributions inside box (Energy Pointing Game metric)
  - Model not pressured to optimize uniformly

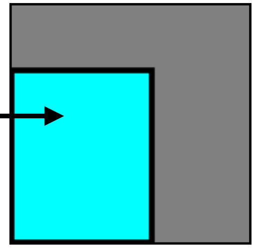
Image

Attribution Map

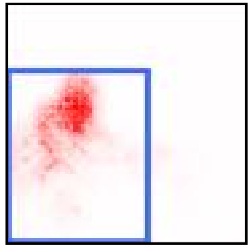


Maximize ***fraction*** of attribution inside Box

Bounding Box



After Guidance

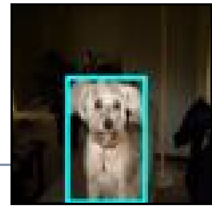


Focus on Object Features

$$\mathcal{L}_{\text{loc},k} = -\frac{\sum_{h=1}^{H}\sum_{w=1}^{W} M_{k,hw} A^{+}_{k,hw}}{\sum_{h=1}^{H}\sum_{w=1}^{W} A^{+}_{k,hw}}$$

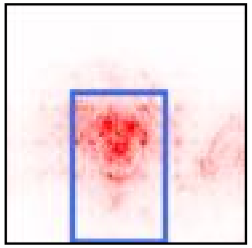
L1 vs. Energy loss

Image



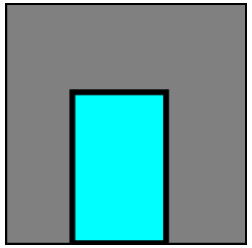
- Popular in prior work: L_1 loss

Attribution Map



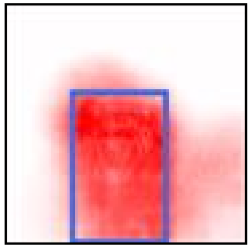
Minimize L_1 distance

Bounding Box



$$\mathcal{L}_{\text{loc},k} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \| M_{k,hw} - \hat{A}^{+}_{k,hw} \|_1$$

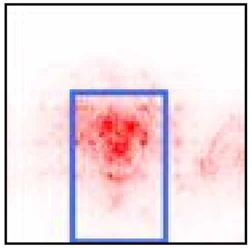
After Guidance



Attributions distributed uniformly in box

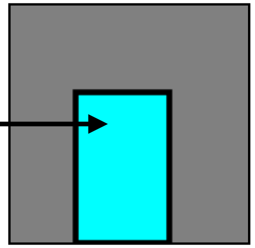
- Our Energy loss

Attribution Map



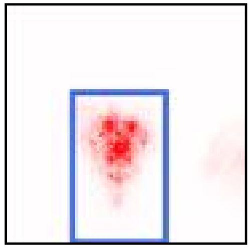
Maximize ***fraction*** of attribution inside Box

Bounding Box



$$\mathcal{L}_{\text{loc},k} = -\frac{\sum_{h=1}^{H} \sum_{w=1}^{W} M_{k,hw} A^{+}_{k,hw}}{\sum_{h=1}^{H} \sum_{w=1}^{W} A^{+}_{k,hw}}$$

After Guidance



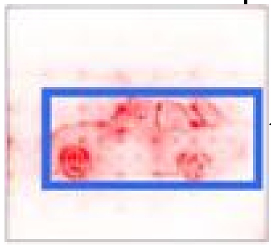
Focus on object features

L1 vs. Energy loss

Image

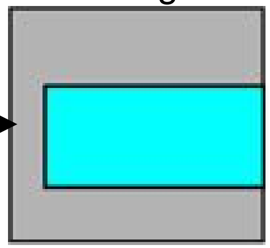
- Popular in prior work: L_1 loss

Attribution Map

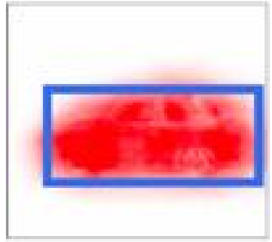


Minimize L_1 distance

Bounding Box



After Guidance

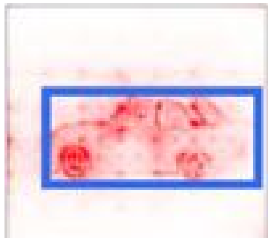


$$\mathcal{L}_{\mathrm{loc},k} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \| M_{k,hw} - \hat{A}^{+}_{k,hw} \|_1$$

Attributions distributed uniformly in box

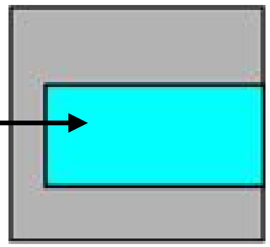
- Our Energy loss

Attribution Map

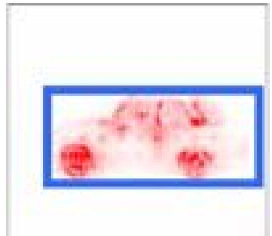


Maximize ***fraction*** of attribution inside Box

Bounding Box



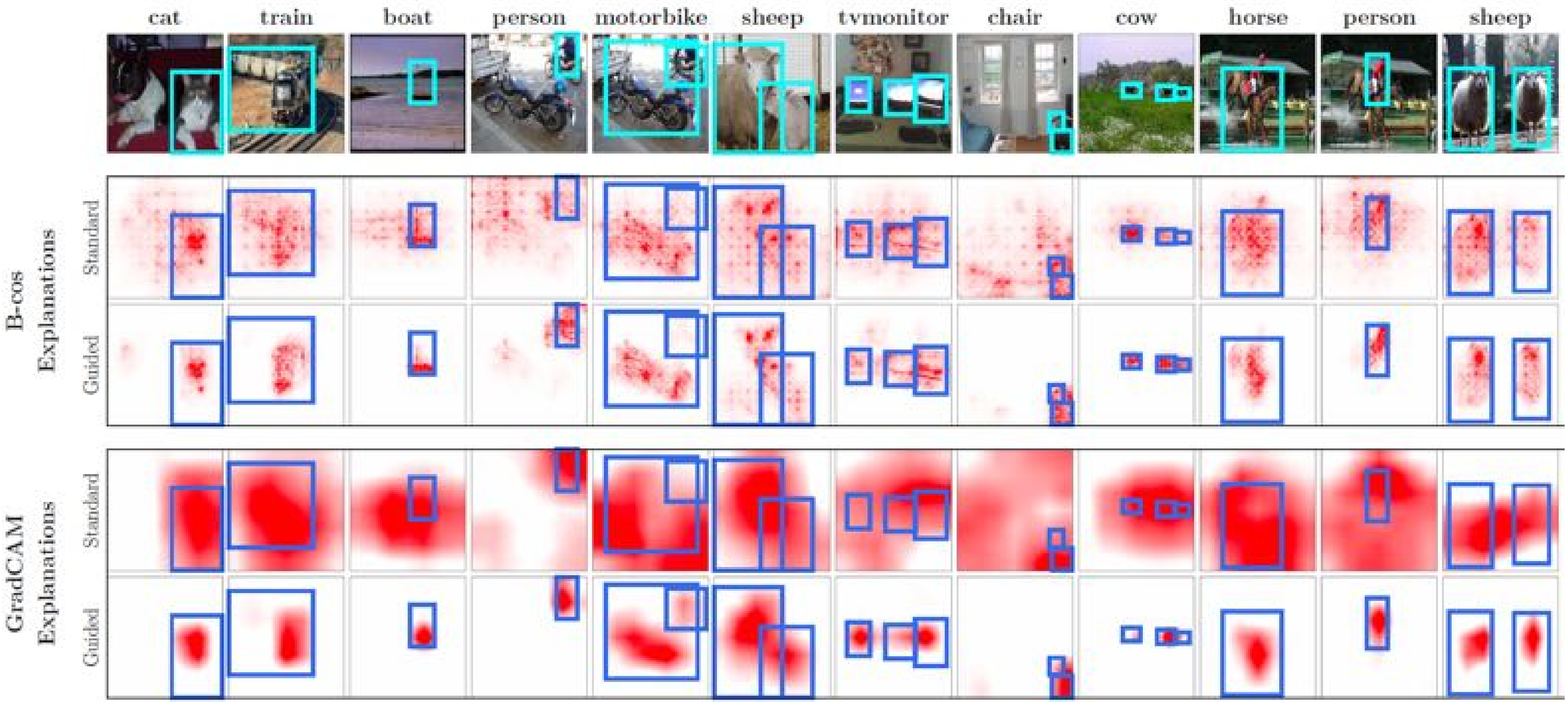
After Guidance



$$\mathcal{L}_{\mathrm{loc},k} = -\frac{\sum_{h=1}^{H} \sum_{w=1}^{W} M_{k,hw} A^{+}_{k,hw}}{\sum_{h=1}^{H} \sum_{w=1}^{W} A^{+}_{k,hw}}$$

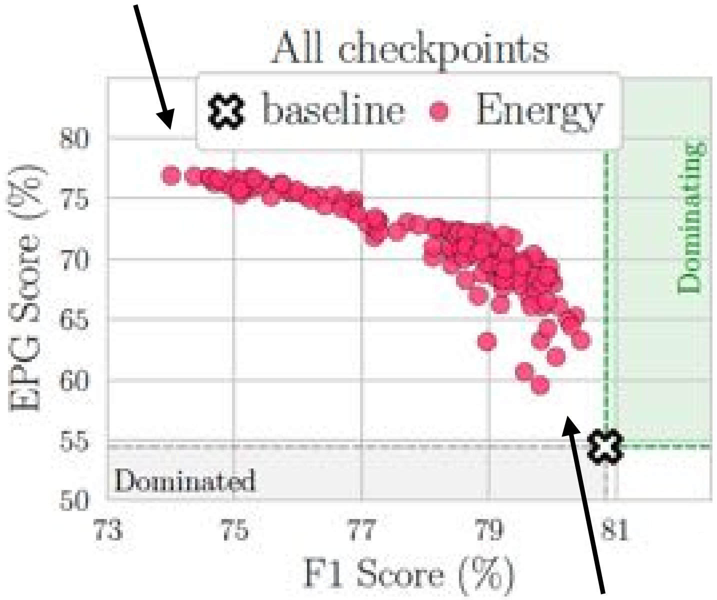
Focus on object features

Qualitative Results



Preliminaries: Visualizing Pareto Fronts

Best localisation,
worst classification

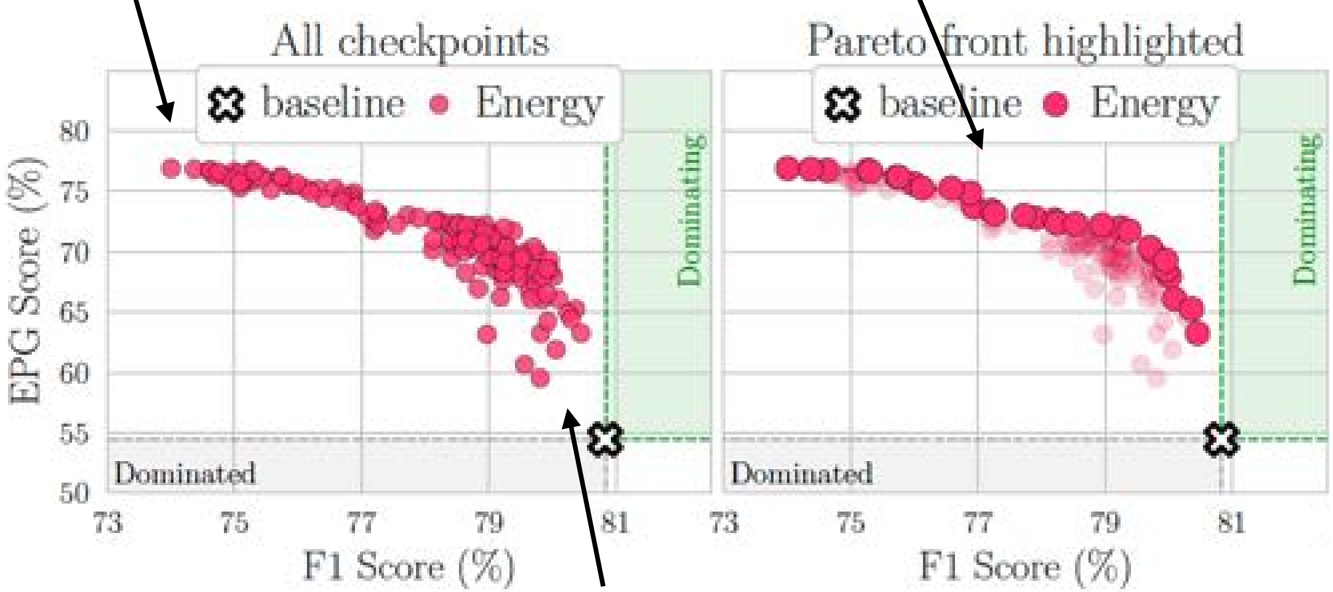


Best classification,
worst localisation

Preliminaries: Visualizing Pareto Fronts

Best localisation,
worst classification

Pareto-dominant
Models



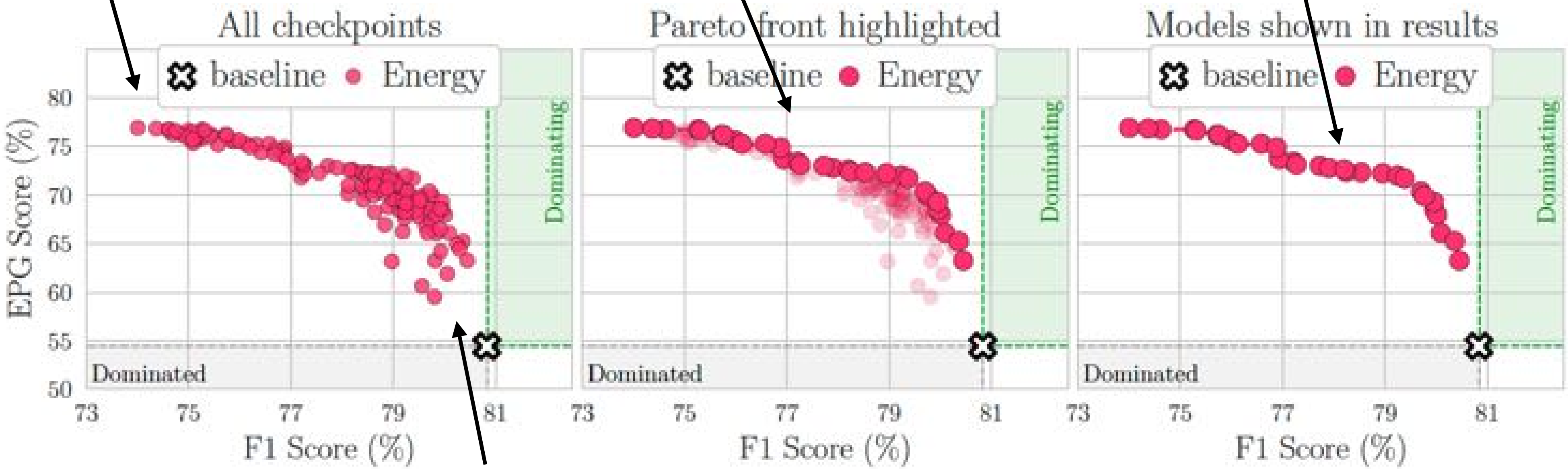
Best classification,
worst localisation

Quantitative Results — Preliminaries: Visualizing Pareto Fronts

Best localisation, worst classification

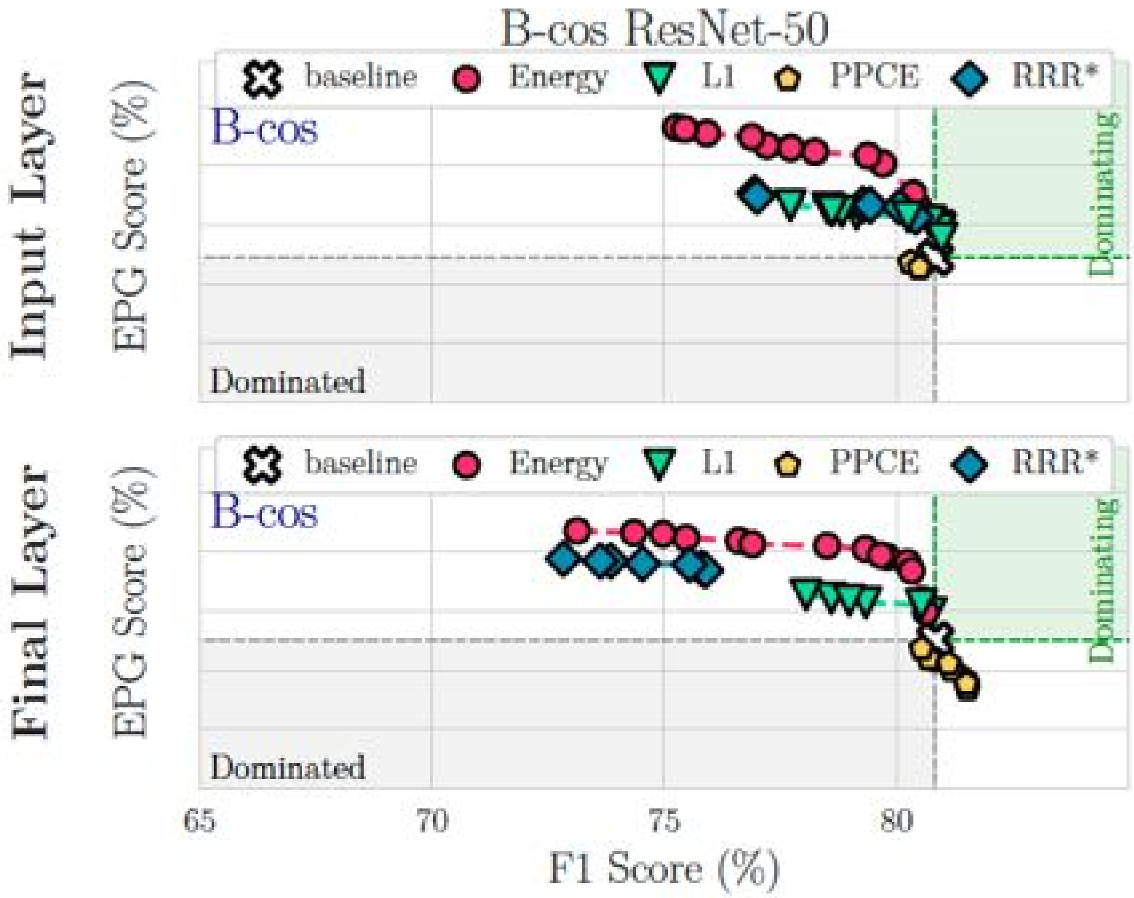
Pareto-dominant Models

We plot and compare models on the Pareto front



Best classification, worst localisation

Quantitative Results: PASCAL VOC



Making guidance cost effective

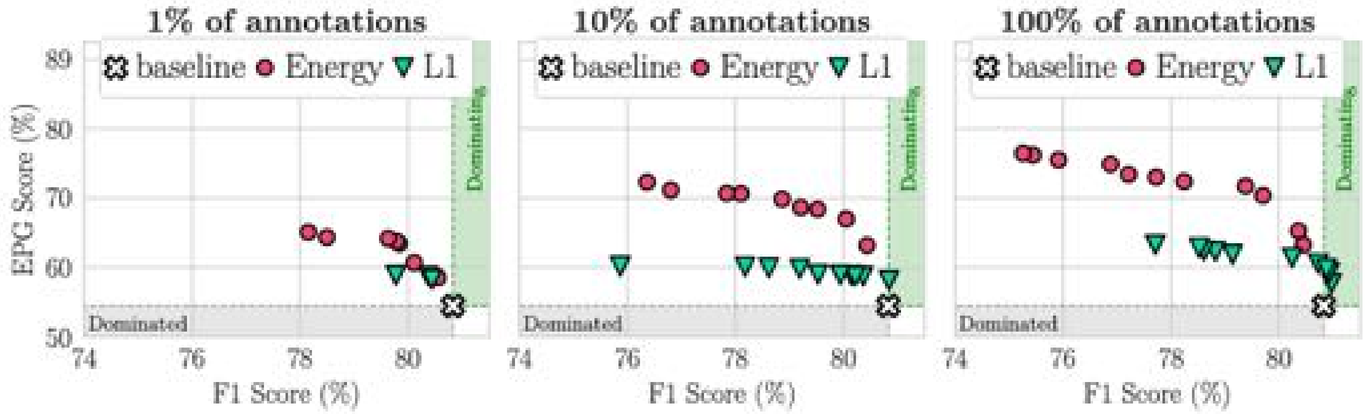
- **So far:**
  - Model guidance helps direct focus on object
  - Energy loss more effectively for focuses on object features as compared to L_1

- **Challenge:**
  - Needs bounding box annotations for a large number of images — costly

- **Reducing annotation cost:**
  - What if we have annotations only for a small fraction of training images?
  - What if bounding box annotations are imprecise and noisy?

Making guidance cost effective

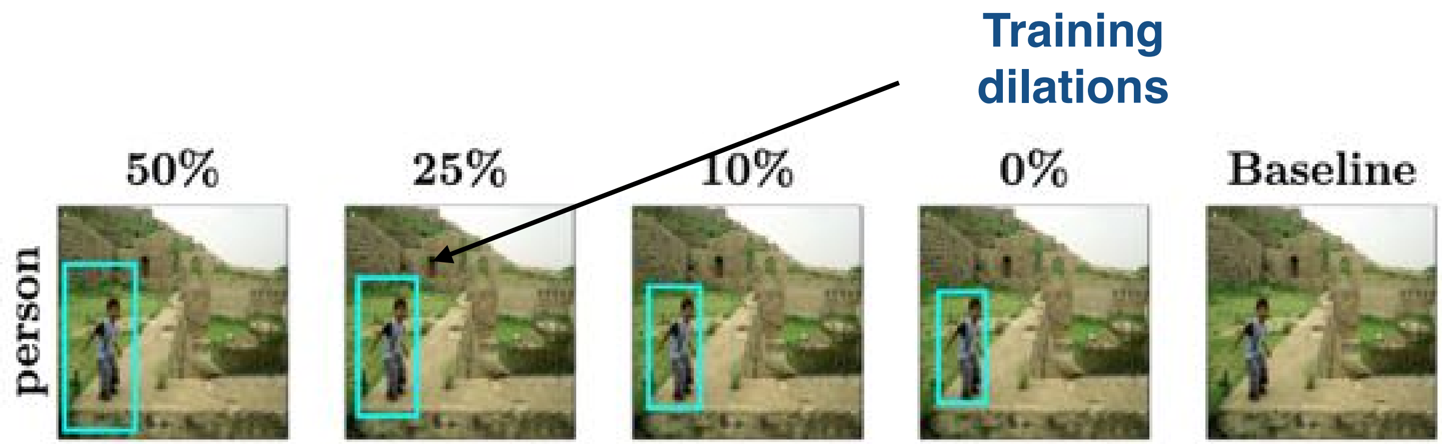
- What if we have annotations only for a small fraction of training images?
- **Experiment:** Use bounding box annotations of only 1% and 10% of training data



- Using 10% annotations performs very similar to using 100% annotations
- Gains even with 1% annotations

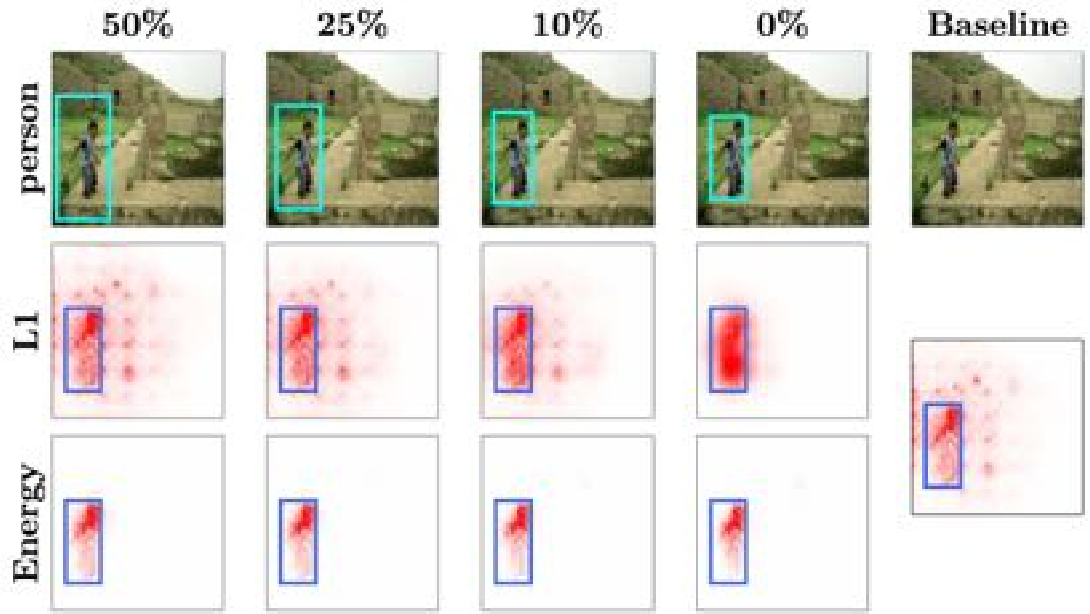
Making guidance cost effective

- What if bounding box annotations are imprecise and noisy? (Easier to annotate)
- **Experiment:**
  - Dilate bounding box to various degrees during training
  - Evaluate with original bounding boxes



Making guidance cost effective

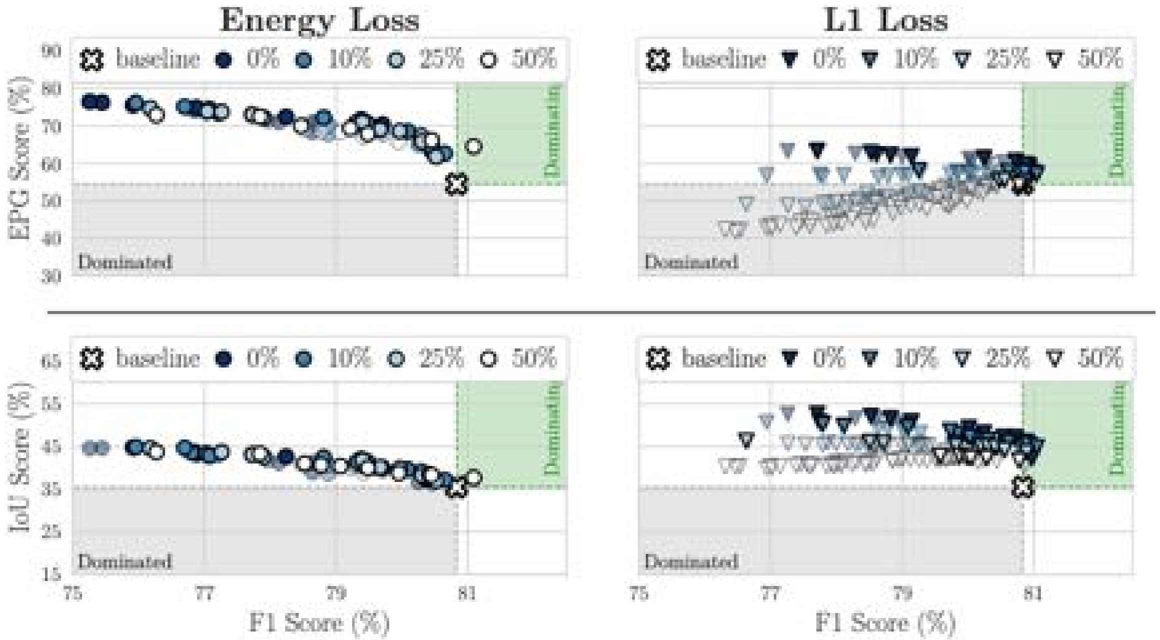
- What if bounding box annotations are imprecise and noisy? (Easier to annotate)



- Energy loss robust, localisation worsens with L_1 loss

Making guidance cost effective

- What if bounding box annotations are imprecise and noisy? (Easier to annotate)



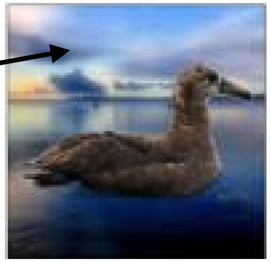
- Energy loss robust, localisation worsens with L_1 loss

Effectiveness against spurious correlations

- **Experiment:** Waterbirds-100, synthetically constructed
- **Training Data:**

Landbird on Land

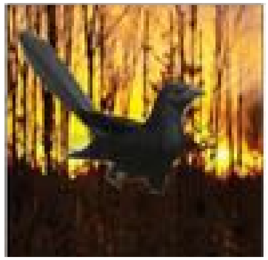
Models often rely on spurious background features



Waterbird on Water

Challenging to classify (Worst group)

- **Test Data:**



Landbird on Land

Landbird on Water

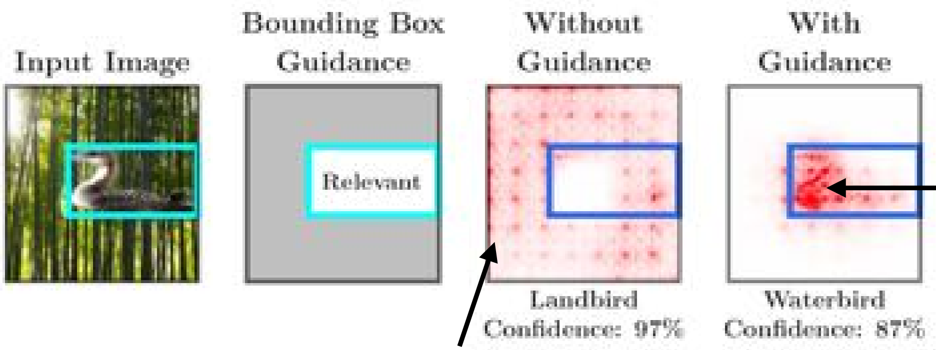
Waterbird on Land

Waterbird on Water

Effectiveness against spurious correlations

- Model guidance shifts focus to the object features, improves accuracy

Waterbird on Land

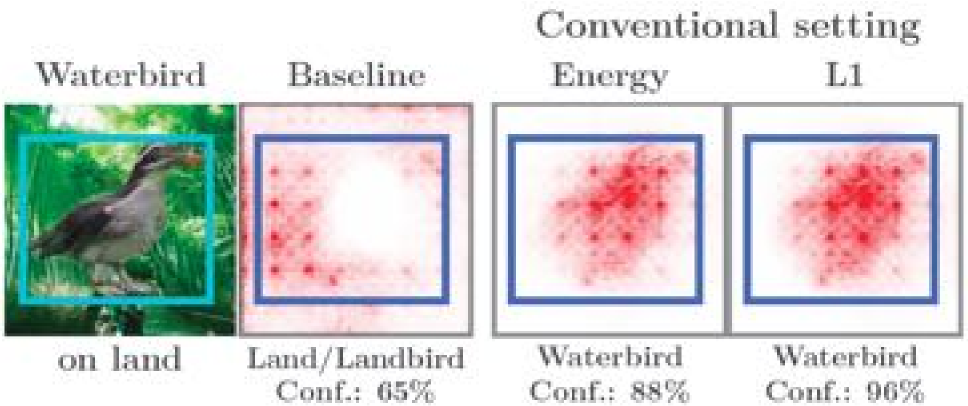


Correct after redirected to focus on object

Incorrect due to reliance on spurious features

Effectiveness against spurious correlations

- Guidance can control whether to focus on the foreground or background



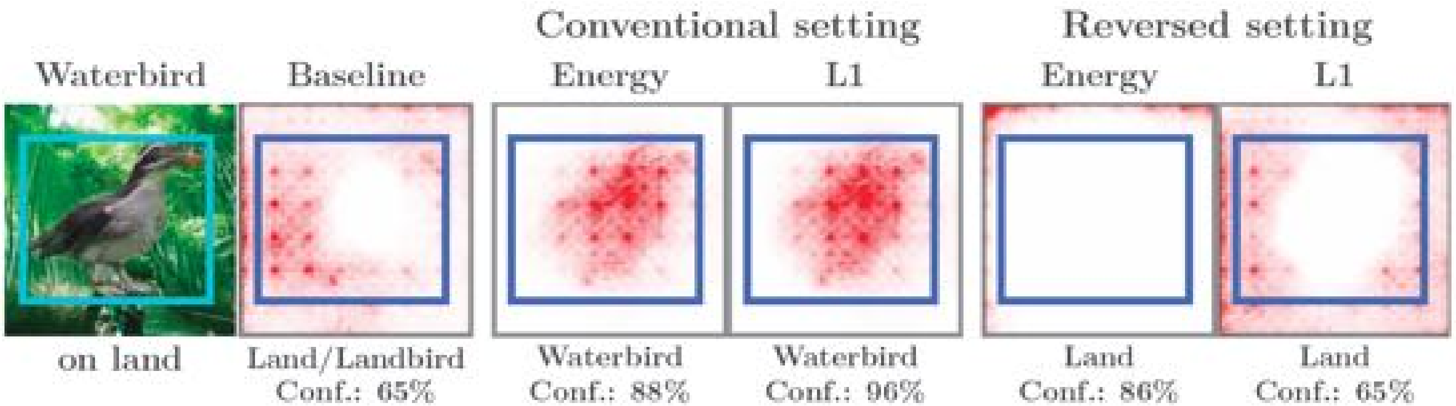
Conventional Setting: direct Focus to the Foreground

	Conventional	
Model	Worst	Overall
Baseline	43.4 (±2.4)	68.7 (±0.2)
Energy	**56.1 (±4.0)**	**71.2 (±0.1)**
L_1	51.1 (±1.9)	69.5 (±0.2)

Guidance improves accuracy

Effectiveness against spurious correlations

- Guidance can control whether to focus on the foreground or background



Reversed Setting: direct Focus to Background

	Conventional		Reversed	
Model	Worst	Overall	Worst	Overall
Baseline	43.4 (±2.4)	68.7 (±0.2)	56.6 (±2.4)	80.1 (±0.2)
Energy	**56.1 (±4.0)**	**71.2 (±0.1)**	**62.8 (±2.1)**	**83.6 (±1.1)**
L_1	51.1 (±1.9)	69.5 (±0.2)	58.8 (±5.0)	82.2 (±0.9)

Guidance improves accuracy

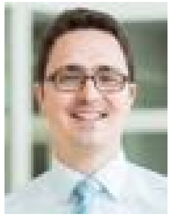
Summary

- **Problem:**
  - Models may reason incorrectly even if they perform well
  - Model guidance can help, but so far not fully explored
- **Contributions:**
  - Propose novel Energy loss
  - Perform comprehensive evaluation on large datasets
  - Show robustness and efficiency of approach
  - Show utility against spurious correlations
- **Outcomes:**
  - Energy loss effective in improving focus, even on large datasets
  - Works with noisy or limited annotations
  - Can improve model performance

Interpretability for Deep Learning in Computer Vision

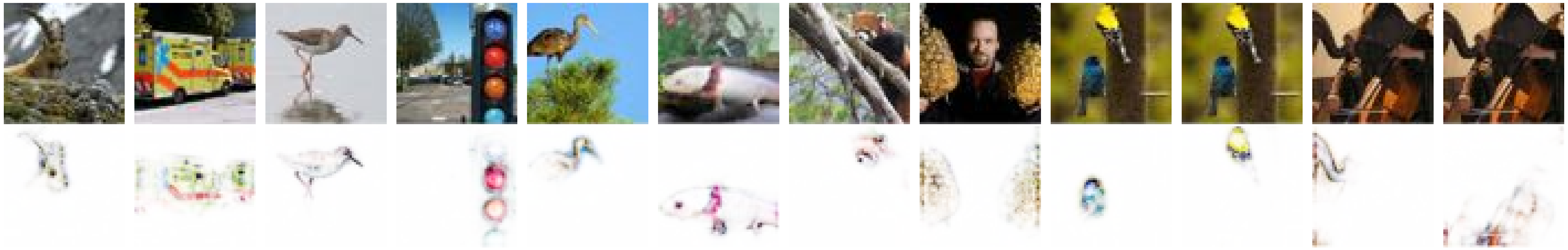
Moritz Boehle

Sukrut Rao



Mario Fritz
CISPA Helmholtz

Bernt Schiele



Max Planck Institute for Informatics & Saarland University,
Saarland Informatics Campus Saarbrücken