Challenges for Deep Learning in Computer Vision:

Interpretability, Robustness and Security

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

Overview

- **Interpretability, Robustness** and **Security** of Deep Learning in Computer Vision
  - **Inherently Interpretable** Deep Neural networks — CVPR’21, CVPR’22
  - **Robustness** of Deep Models:
    Bright and **Dark** Side of **Scene Context —** NeurIPS'18, CVPR'19, ECCV'20
  - **Security** of Deep Models
    Reverse Engineering and **Stealing** of Deep Models — ICLR’18, CVPR’19, ICLR’20

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 Interpretabilituy

@ CVPR 2022

Moritz Boehle
MPI Informatics

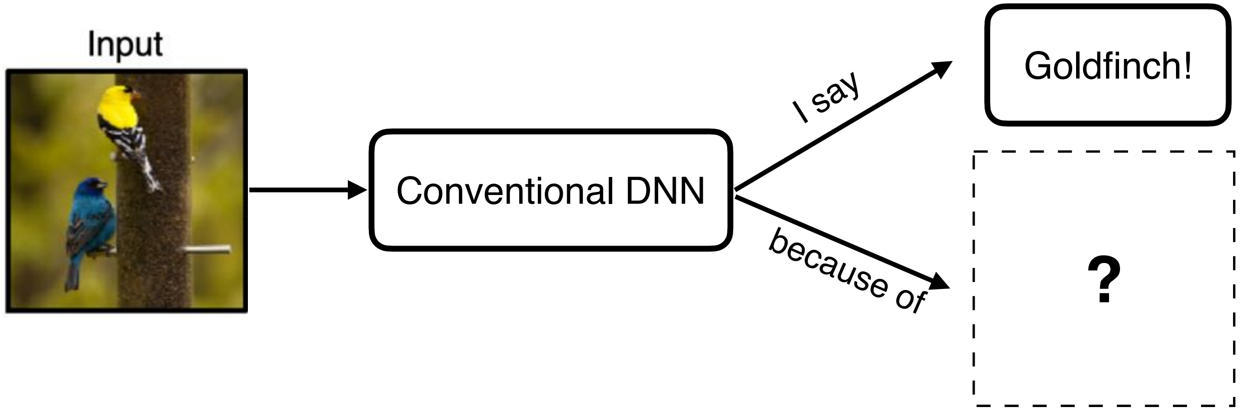
Mario Fritz
CISPA Helmholtz

Bernt Schiele
MPI Informatics

Motivation

Want | Explanation | to | understand better

Requirements | Faithful | Interpretable

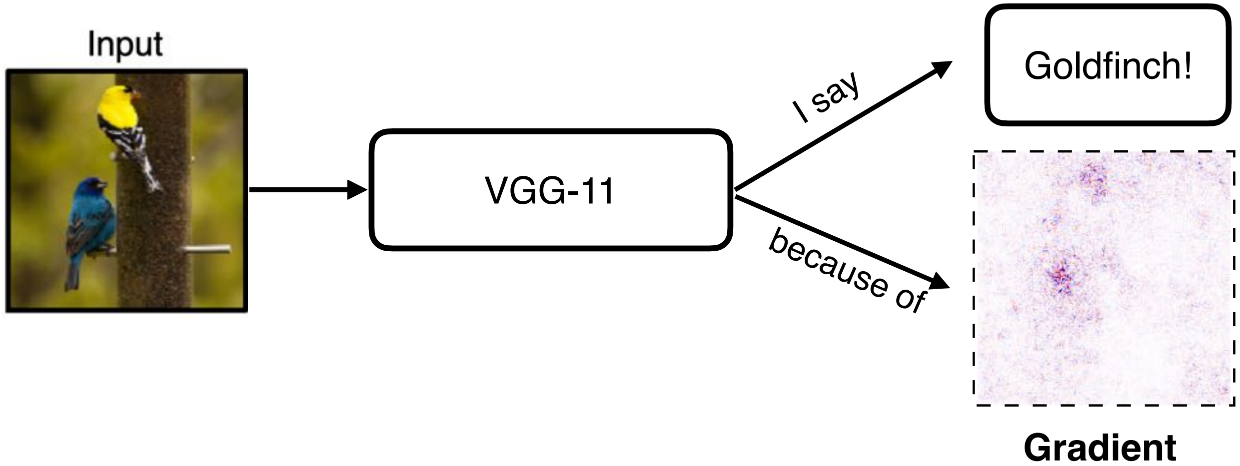


References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

Want Explanation

Requirements Faithful ✓ Interpretable ✗



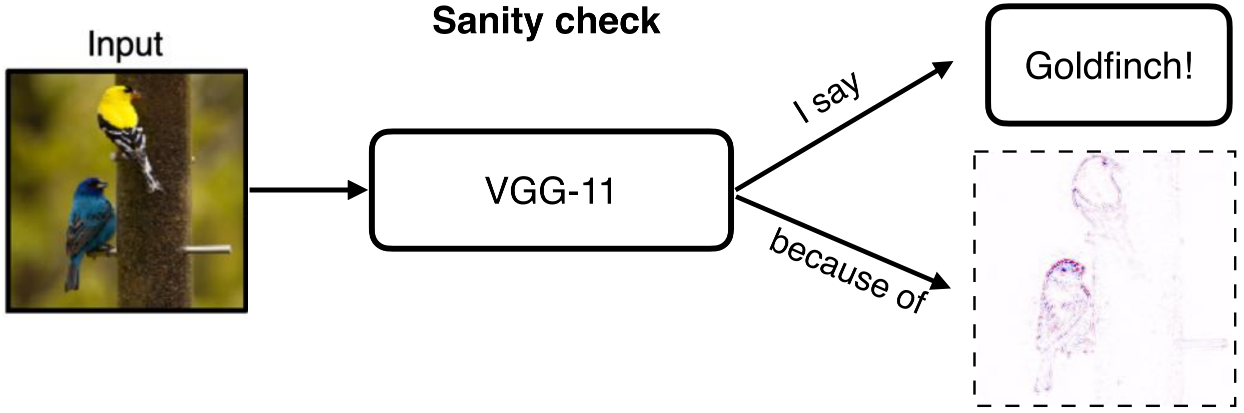
References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

Want: Explanation

Requirements: Faithful | Interpretable ✓

Sanity check



Guided Backpropagation

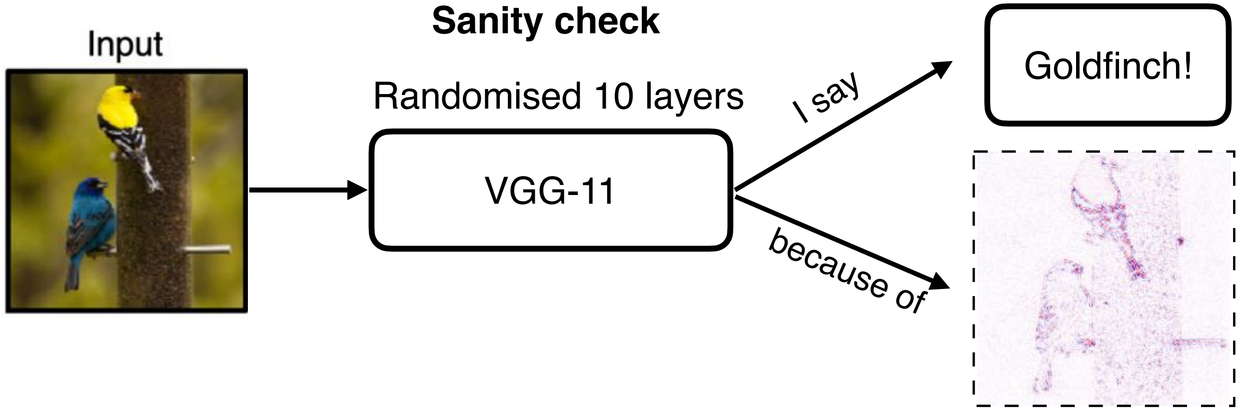
References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

Want: Explanation

Requirements: Faithful ✗ | Interpretable ✓

Sanity check



Guided Backpropagation

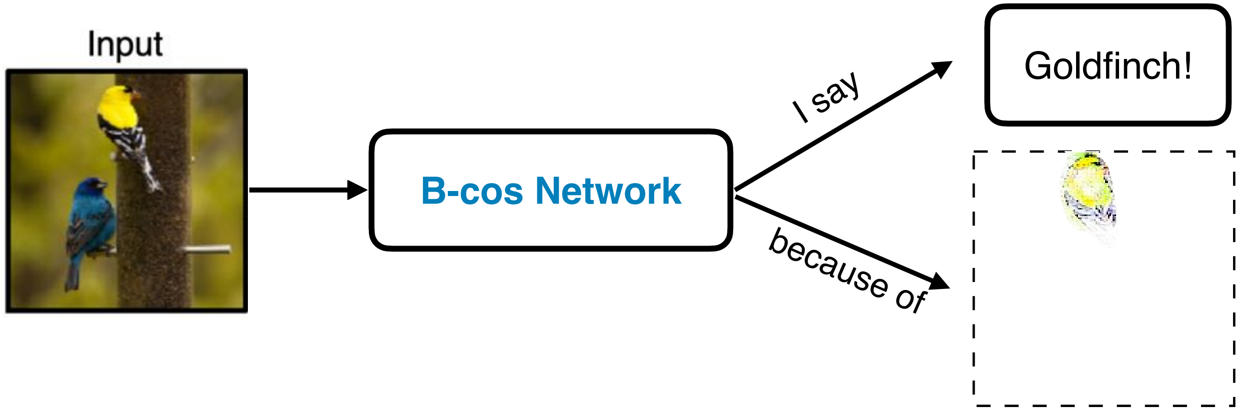
References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation: we aim for Inherent Interpretability

Want — Explanation

Requirements — Faithful ✓ — Interpretable ✓

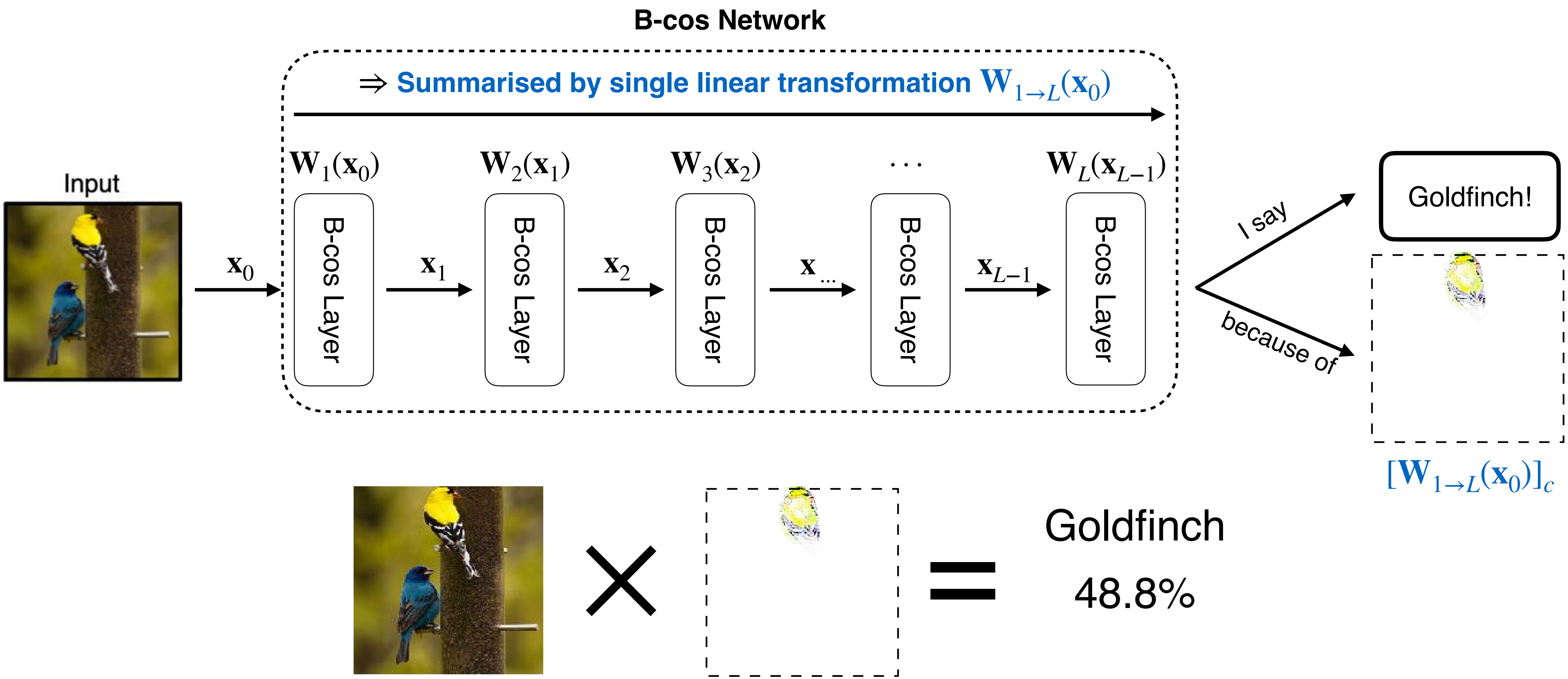
Dynamic linearity — **Alignment pressure**



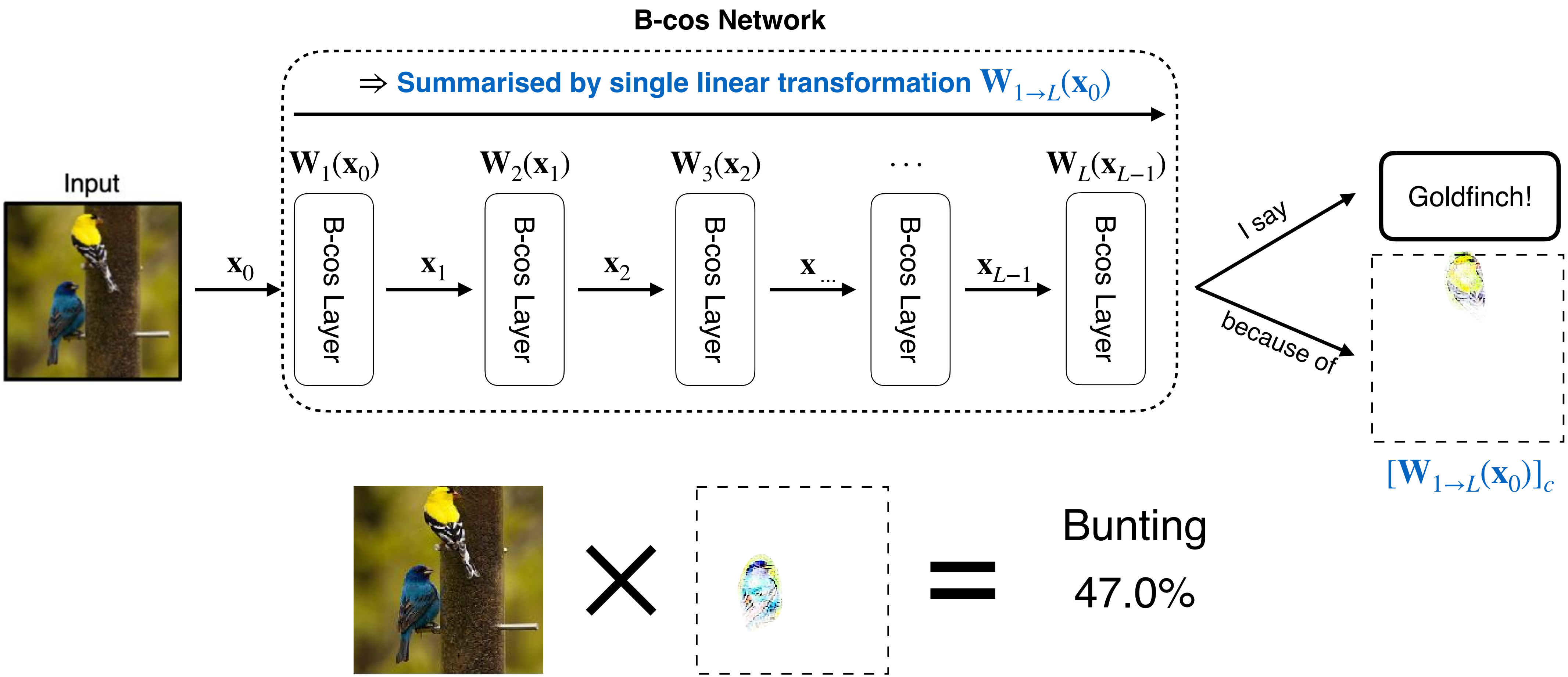
Model-inherent linear map

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)

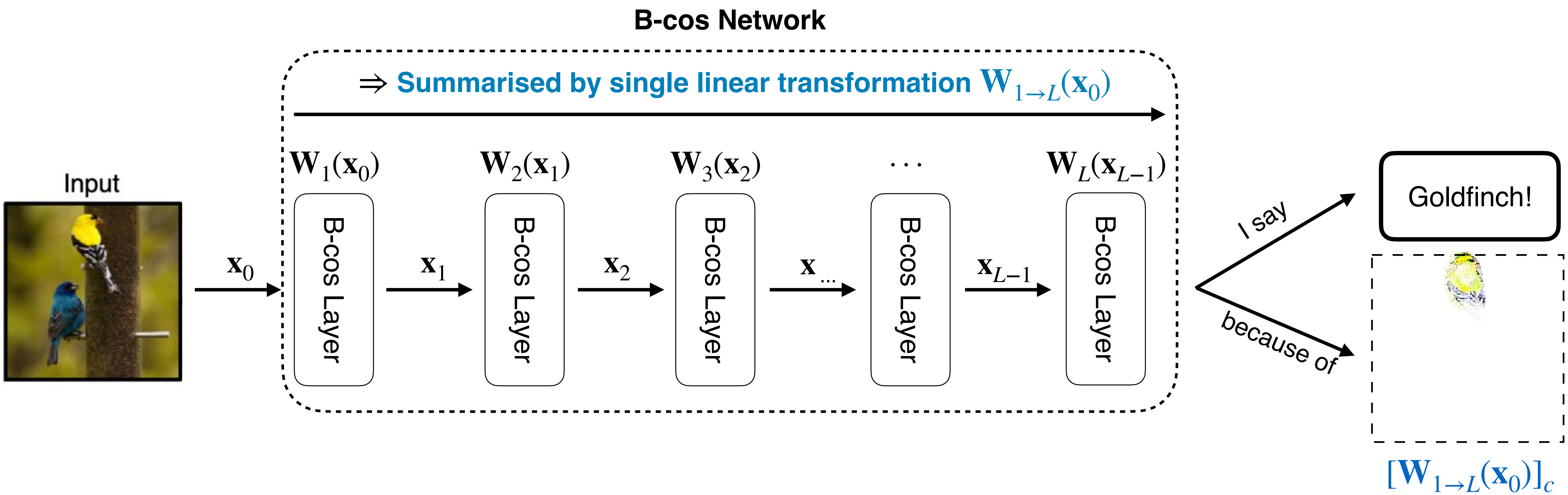
Dynamic linearity



Dynamic linearity



Dynamic linearity



Dynamic linearity allows us to faithfully summarise the model.

max planck institut
informatik

Alignment pressure

B-cos transformation vs. linear transformation

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

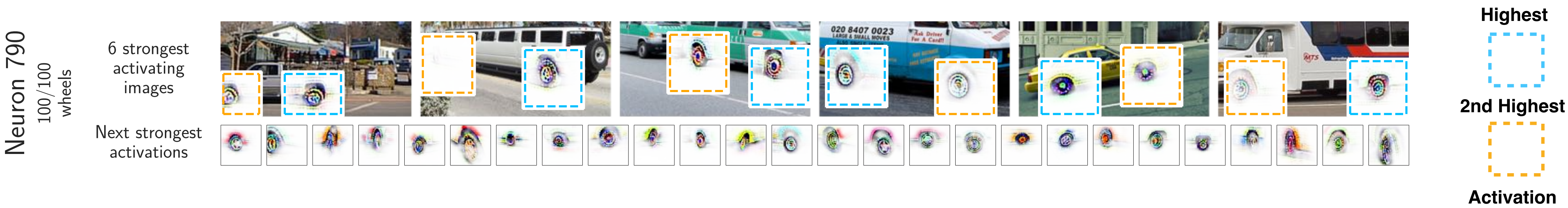
New transformation B-cos$(\mathbf{x};\ \mathbf{w}) = \underbrace{||\,\widehat{\mathbf{w}}\,||}_{=1}\ ||\mathbf{x}||\ |\cos(\mathbf{x}, \mathbf{w})|^{B} \times \mathrm{sgn}\left(\cos(\mathbf{x}, \mathbf{w})\right)$

slide credit: Moritz Böhle

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

slide credit: Moritz Böhle

Visualisations: intermediate neurons



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, Robustness** and **Security** of Deep Learning in Computer Vision
  - **Inherently Interpretable** Deep Neural networks — CVPR’21, CVPR’22
  - **Robustness** of Deep Models:
    Bright and **Dark** Side of **Scene Context** — NeurIPS'18, CVPR'19, ECCV'20
  - **Security** of Deep Models
    Reverse Engineering and Stealing of Deep Models — ICLR’18, CVPR’19, ICLR’20

max planck institut informatik

SIC Saarland Informatics Campus

Adversarial Scene Editing: Automatic Object Removal from Weak Supervision

@ NeurIPS 2018

Not Using the Car to See the Sidewalk: Quantifying and Controlling the Effects of Context in Classification and Segmentation

@ CVPR 2019

Rakshith Shetty
MPI Informatics

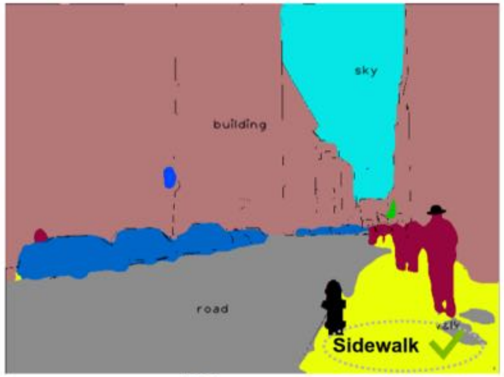
Mario Fritz
CISPA Helmholtz

Bernt Schiele
MPI Informatics

Motivation: The Bright and the Dark Side of Scene Context

- Current models heavily rely on scene context:
  - Original image with cars on the left side:

original $(\mathcal{I})$



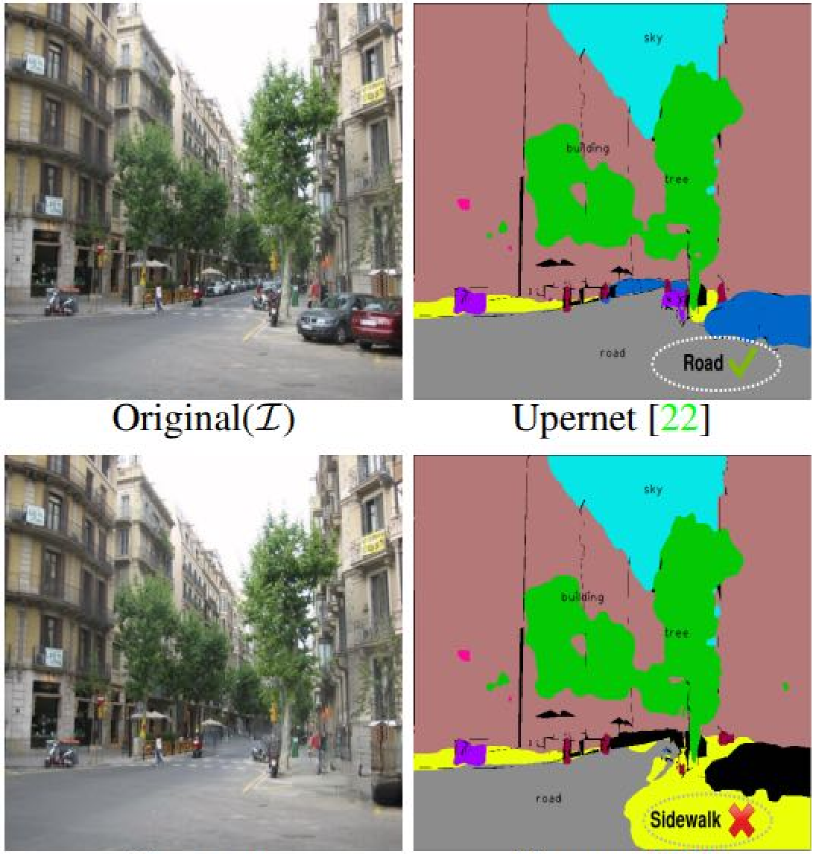
Upernet

  - Same image without those cars:

Question: How Dependent are Current Models on Scene Context?

- Here
  - we look at a **particular aspect** of **context** : co-occurring objects

- Goals:
  - **quantify context sensitivity** of classification and segmentation using **object removal** [NeurIPS'18]
  - object removal based **data augmentation** for **better performance**

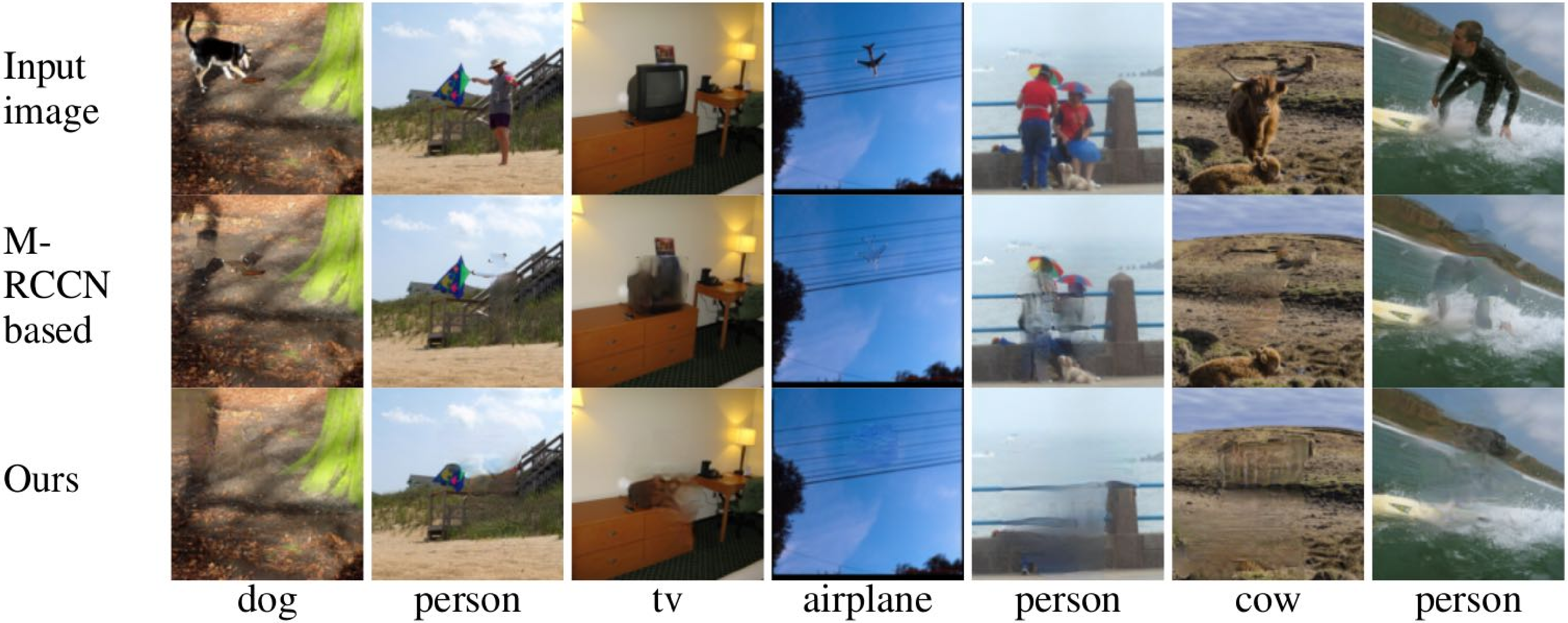


Original($\mathcal{I}$) Upernet [22]

$\mathcal{I} - car$ Upernet [22]

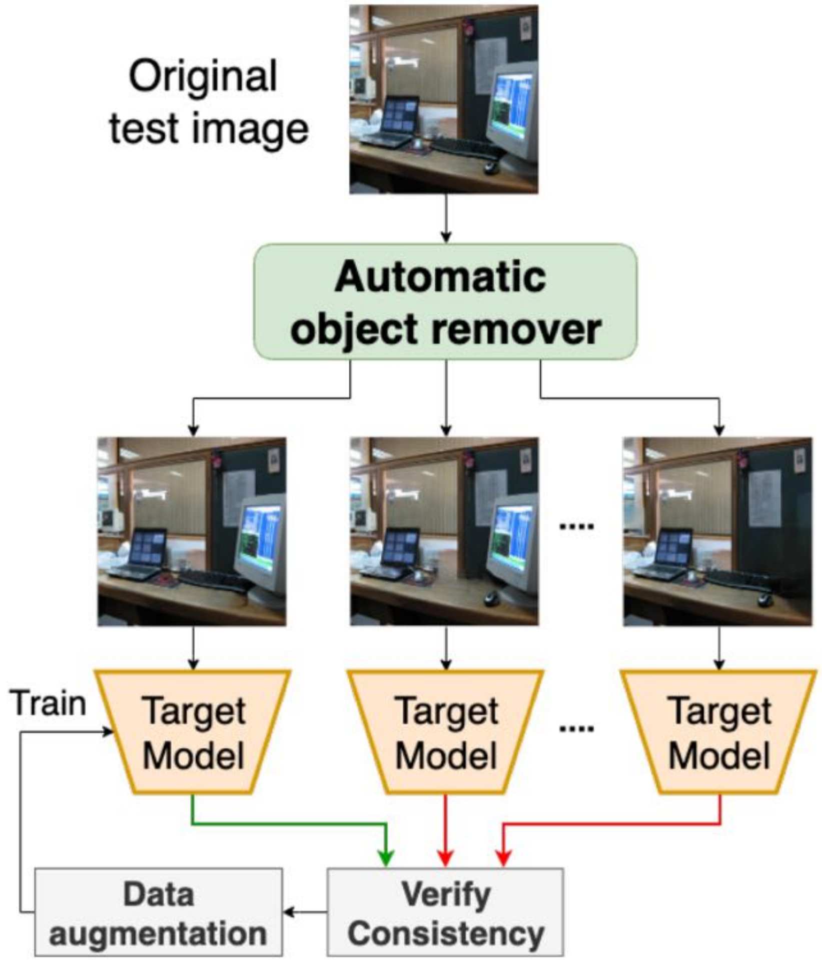
[Shetty, Fritz, Schiele, NeurIPS'18]

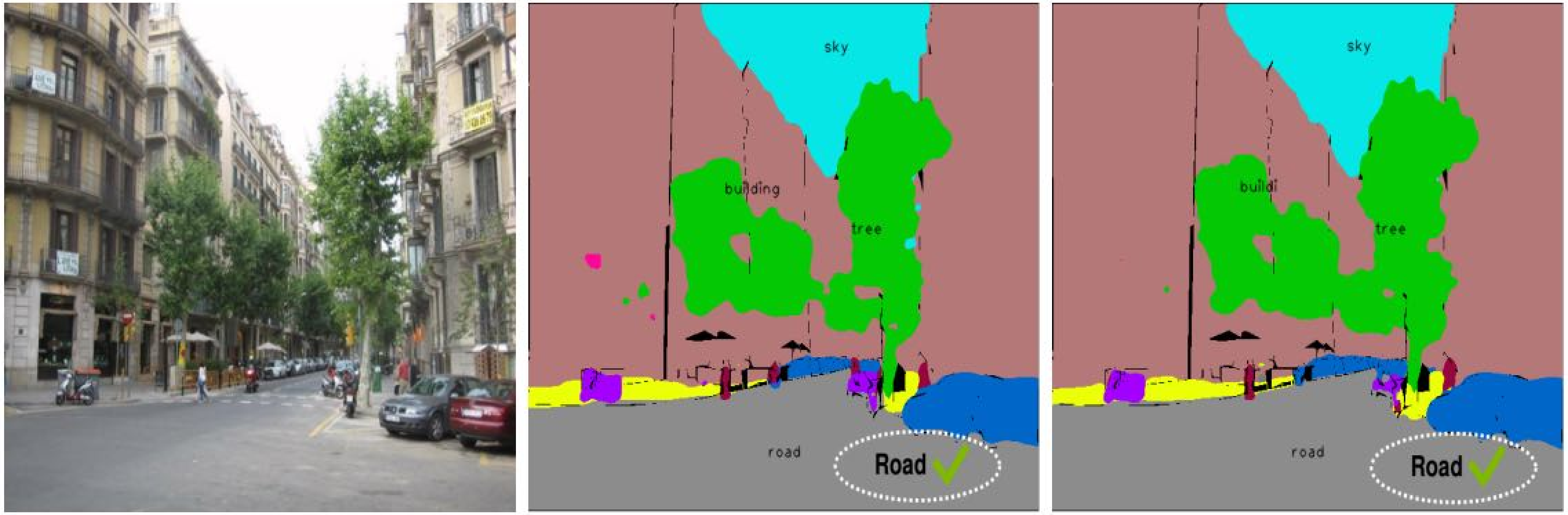
Qualitative Results - COCO Dataset



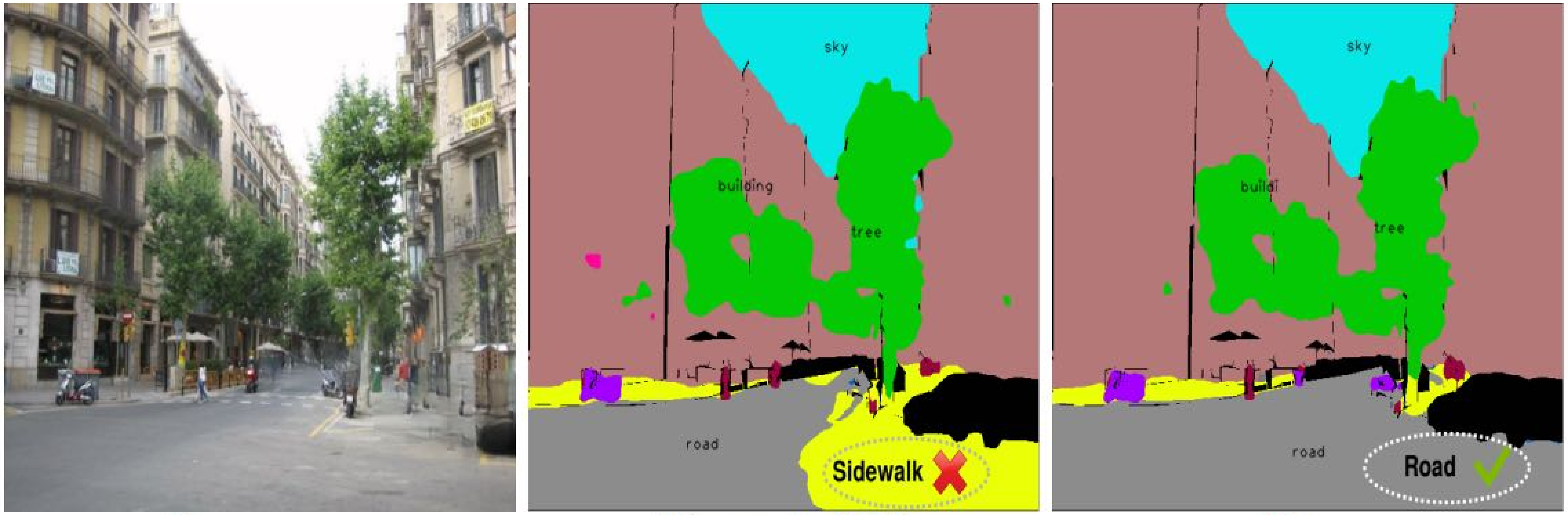
Automated Testing Framework

- Idea:
  - create multiple versions of the input image with one obje
- Removal approach: [Shetty, Fritz, Schiele, NeurIPS'18]
  - use ground truth masks + in-painter trained for object re
- Each image presents new context in the “neighbo

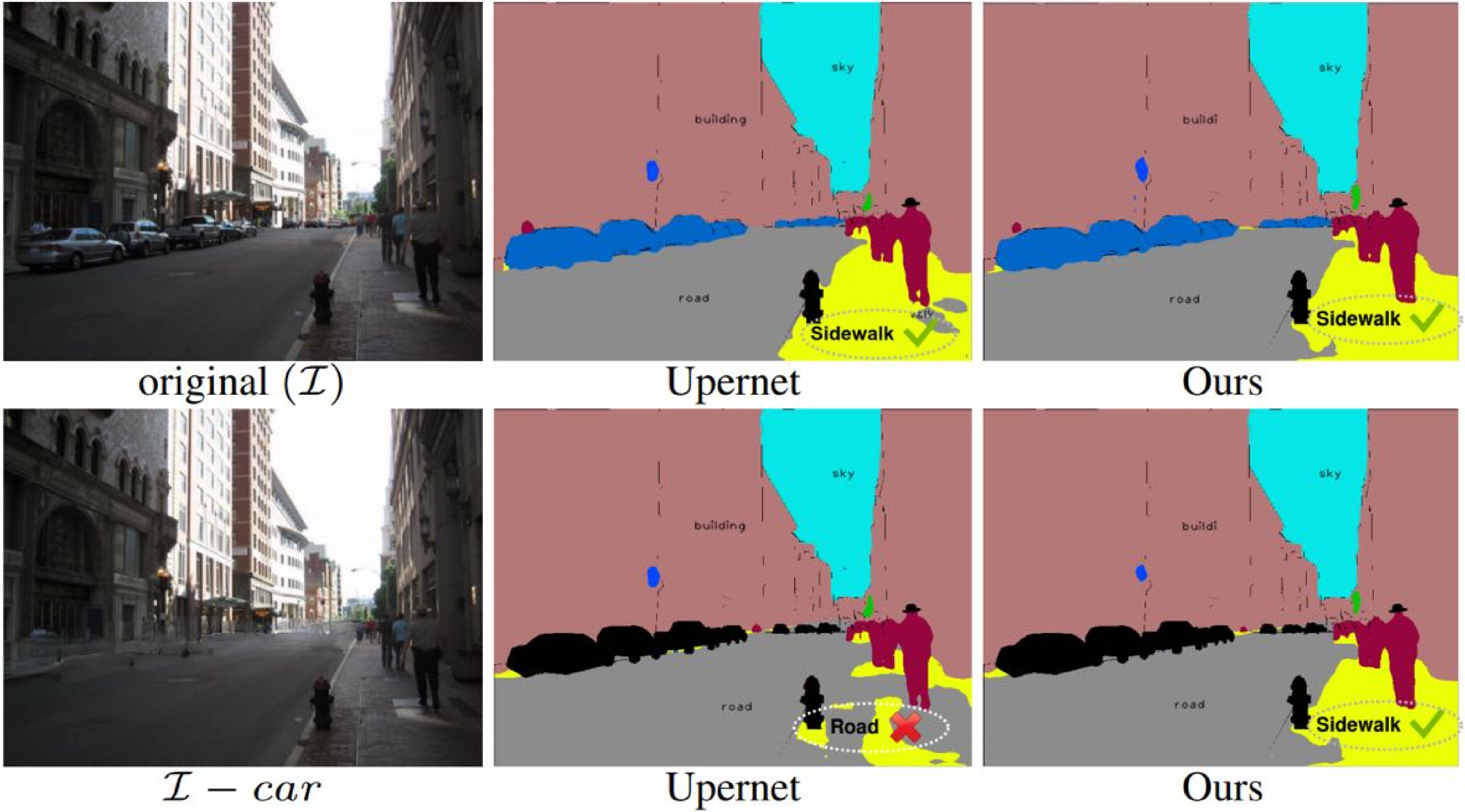


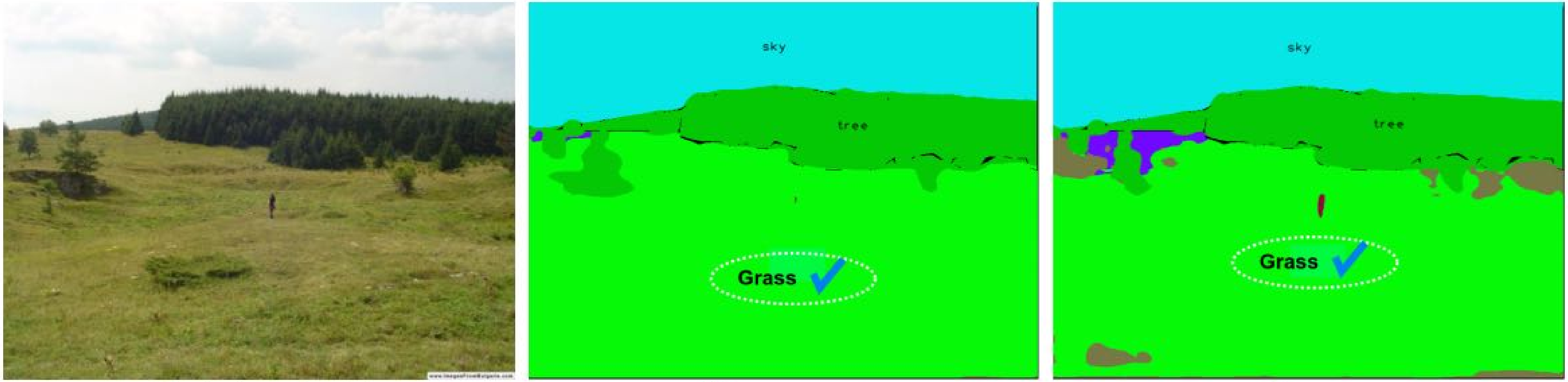


Original($\mathcal{I}$) Upernet [22] Ours

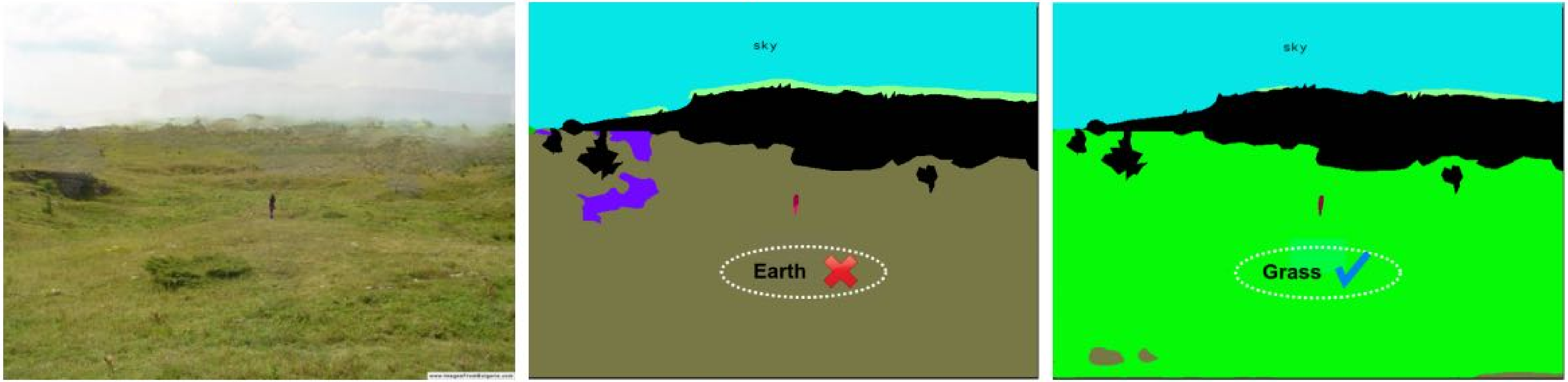


$\mathcal{I} - car$ Upernet [22] Ours

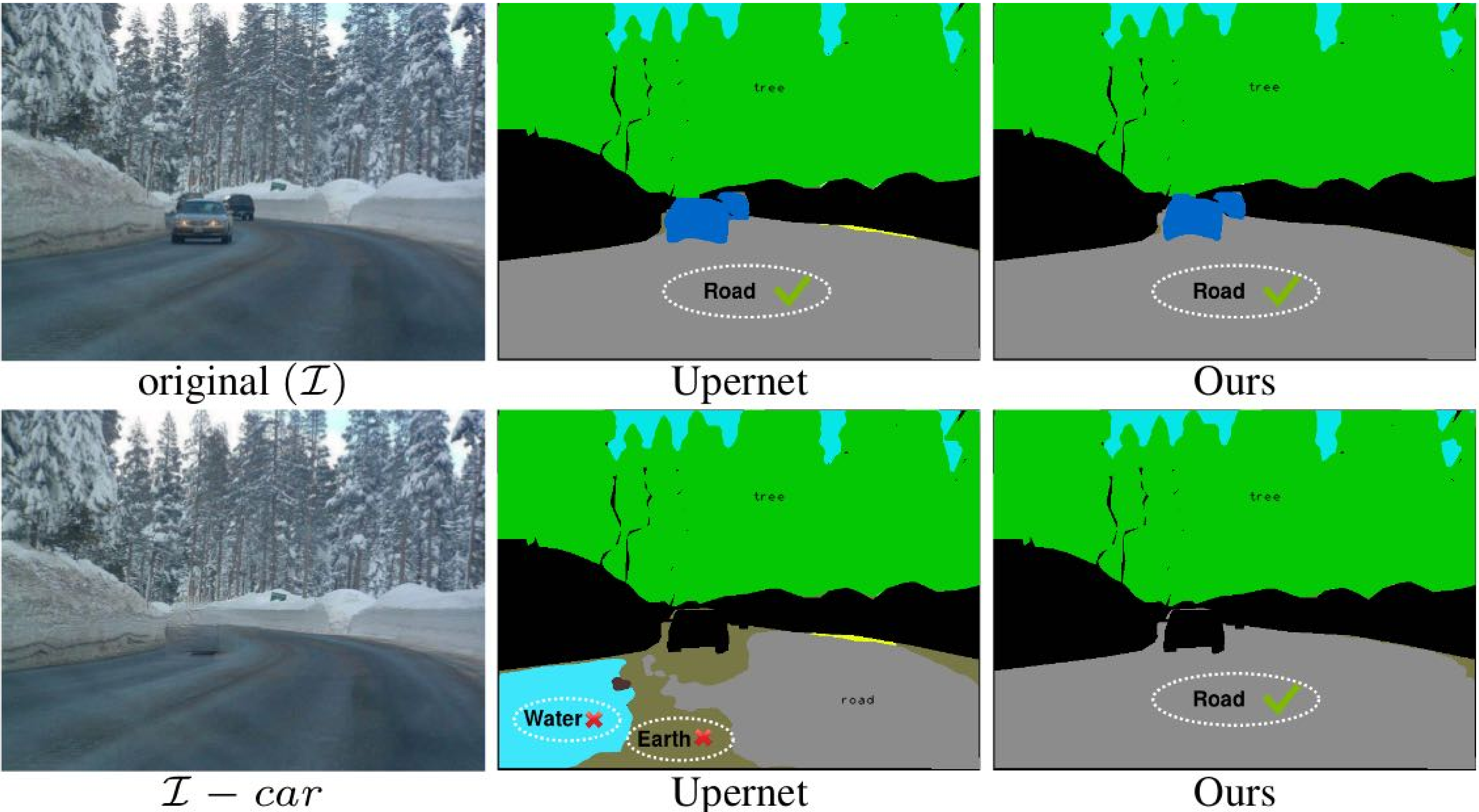

sky
building
road
Sidewalk
buildi
Road
original ($\mathcal{I}$)
Upernet
Ours
$\mathcal{I} - car$
Upernet
Ours



Original: $\mathcal{I}$ Upernet Ours



Edited: $\mathcal{I} - tree$ Upernet Ours


tree
Road
Road
original ($\mathcal{I}$)
Upernet
Ours
tree
road
Water
Earth
Road
$\mathcal{I} - car$
Upernet
Ours

max planck institut informatik

Saarland Informatics Campus

Towards Automated Testing and Robustification by Semantic Adversarial Data Generation

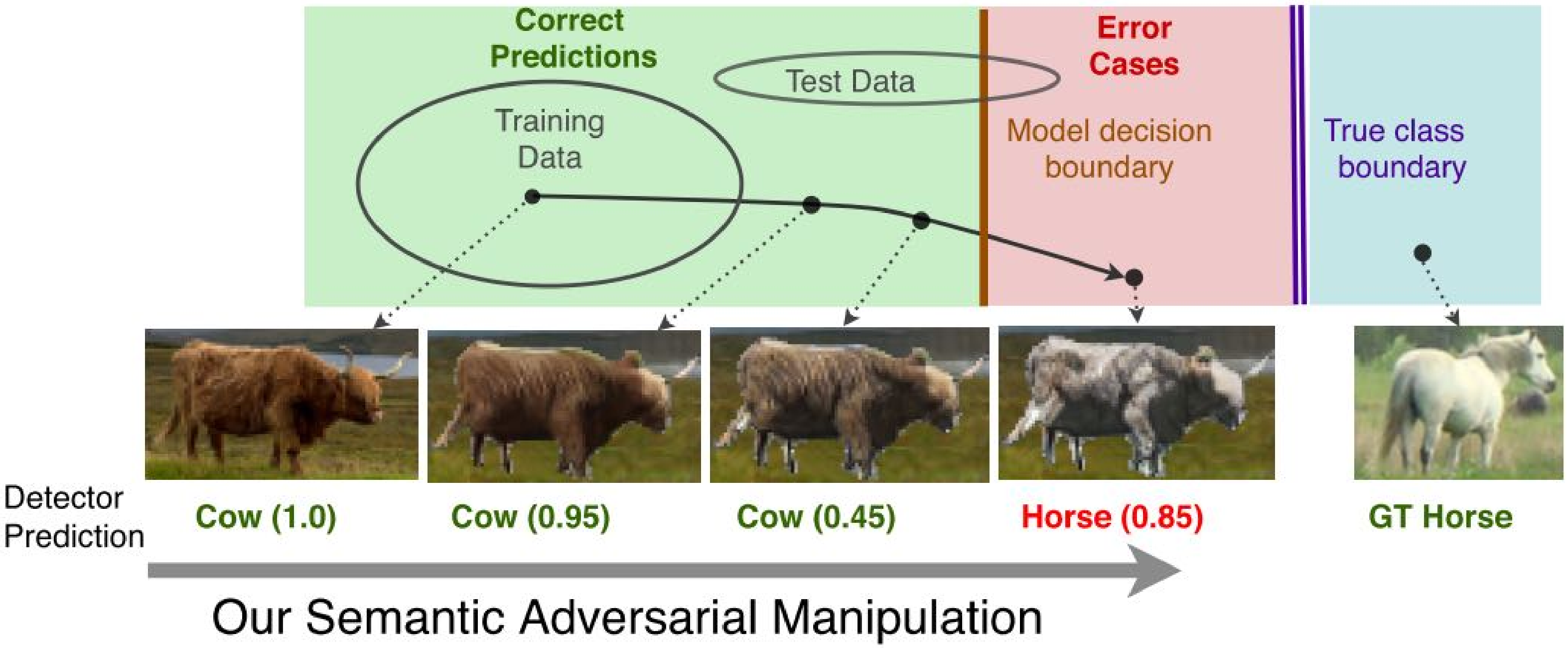
@ ECCV 2020

Rakshith Shetty
MPI Informatics

Mario Fritz
CISPA Helmholtz

Bernt Schiele
MPI Informatics

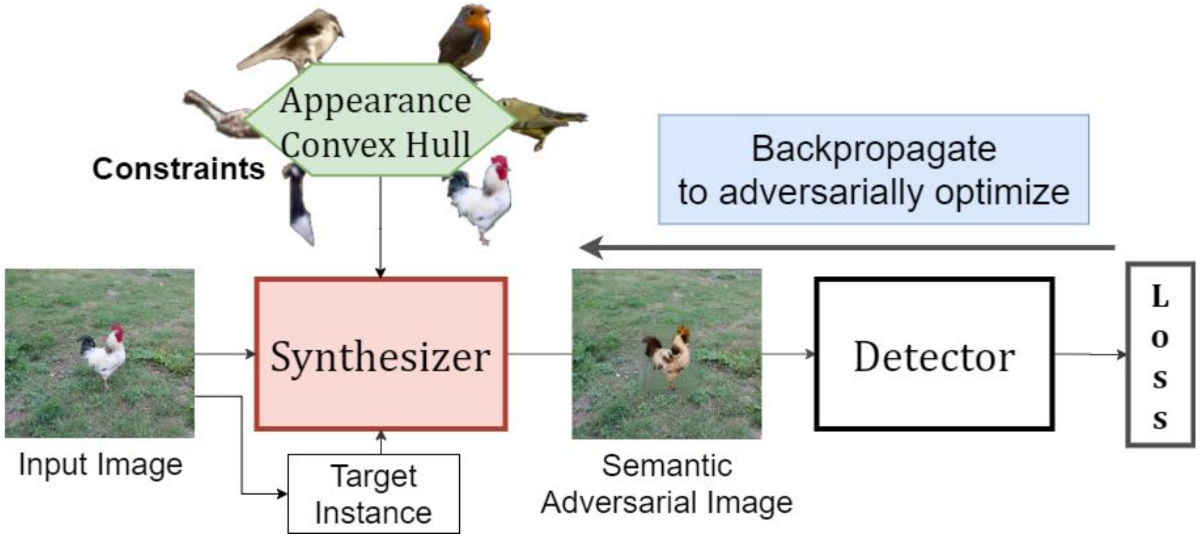
Model adaptive testing via semantic adversary



- Core Idea: Use a **generative model** + constrained **adversarial attack** to move in the data space and synthesize targeted novel failure modes

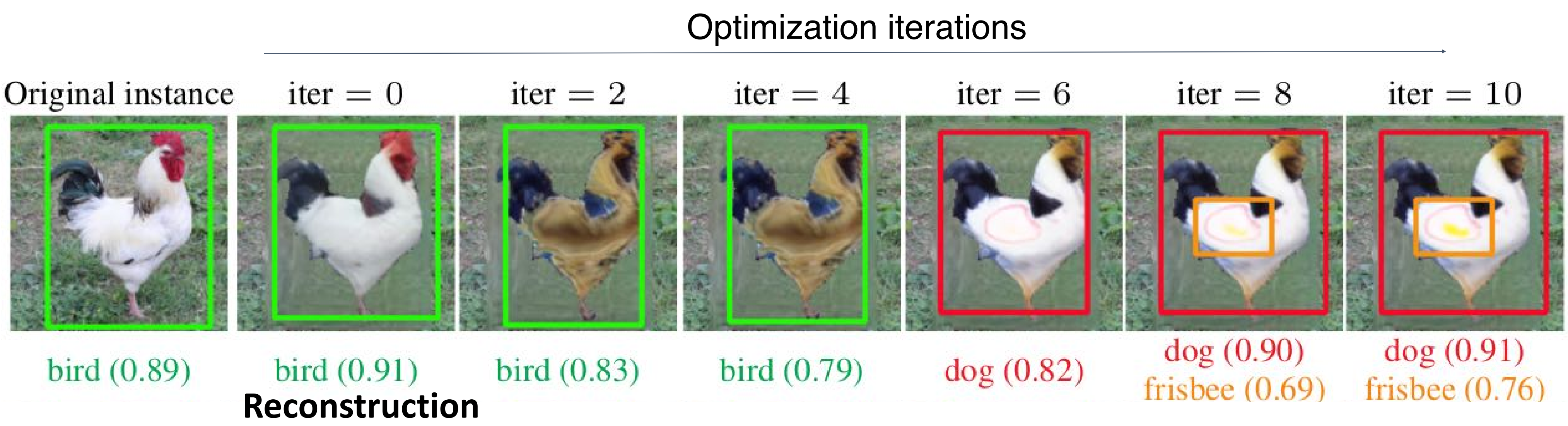
Key components

Details in the ECCV'20 video & paper



- A **synthesizer** → generates objects with disentangled shape and appearance
- **Adversarial optimization** → guide the synthesis to towards hard cases
- **Appearance constraints** → keep synthesized appearance realistic

Illustration of the semantic adversarial attack

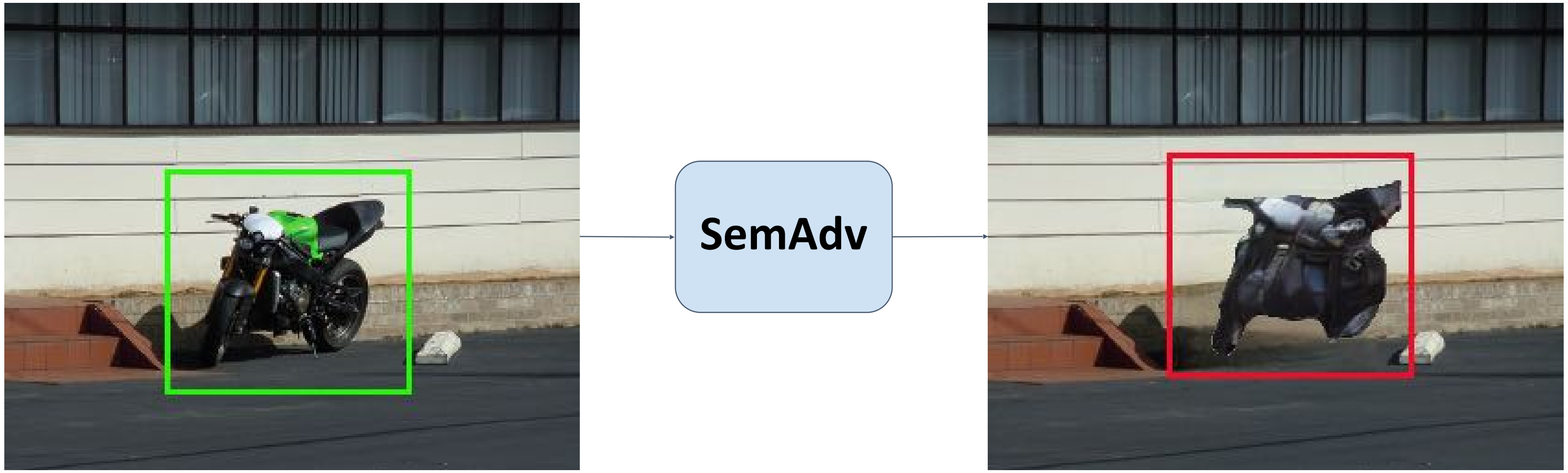


Synthesized hard examples - Camouflaging

SemAdv

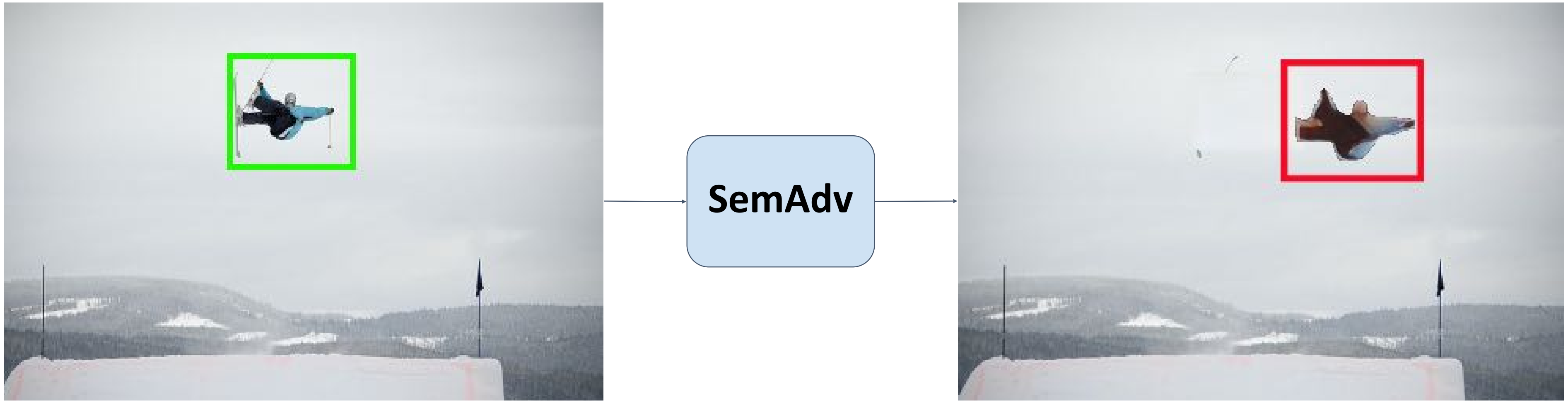
Prediction : **Stop Sign** ✓ **No detection** ✗

Synthesized hard examples - Appearance



Prediction : **Motorcycle** ✓ **Backpack** ✗

Synthesized hard examples - Context



Prediction : **Person ✓** **Airplane ✗**

More examples in the paper & the supplementary

Data augmentation results: Summary

- Small but consistent improvement on **three datasets**
- Larger gains on out-of-dataset distribution test samples

	IID test set	**OOD test set**
COCO	+ 2.17%	**+ 4.6%**
PASCAL VOC	+ 1.35%	**+ 4.9%**
BDD 100k	+ 1.38%	+ 1.15%

Take Home Message - Towards more Robust Models

- The **bright and dark sides of scene context**
  - scene context helps to achieve better performance - however **current models** are **too dependent** on **scene context**
- Proposed **new testing framework** and **data augmentation framework**
  - **automatically generate diverse** set of **scene context** (via object removal)
  - allows to **overcome some** of the **context dependencies**
- Proposed **new semantic adversarial generation framework**
  - generate **"semantically" constrained failure cases beyond i.i.d.**
  - for automated testing and robustification
- **More work required** !

Overview

- **Interpretability, Robustness** and **Security** of Deep Learning in Computer Vision
  - **Inherently Interpretable** Deep Neural networks — CVPR’21, CVPR’22
  - **Robustness** of Deep Models:
    Bright and **Dark** Side of **Scene Context** — NeurIPS'18, CVPR'19, ECCV'20
  - **Security** of Deep Models
    Reverse Engineering and Stealing of Deep Models — ICLR’18, CVPR’19, ICLR’20

Towards Reverse Engineering Black-Box Neural Networks
@ ICLR 2018

Knockoff Net: Stealing Functionality of Black-Box Models
@ CVPR 2019

Prediction Poisoning: Towards Defenses Against DNN Model Stealing Attacks
@ ICLR 2020

Tribhuvanesh Orekondy
MPI Informatics

Seong Joon Oh
MPI Informatics

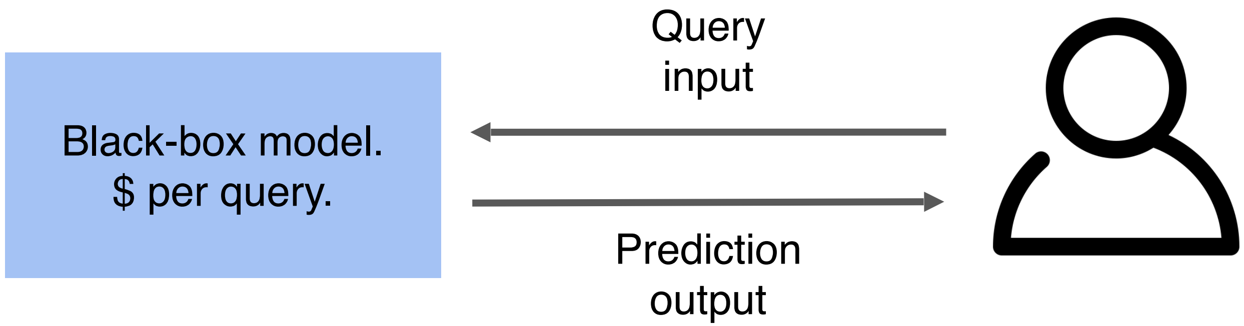
Bernt Schiele
MPI Informatics

Mario Fritz
CISPA Helmholtz

Providing ML Models is a Business Model

- **Input in, prediction out**. Ask $ per query.
  - ML models are **black boxes** !
  - not shared: **architecture**, **parameters**, **hyperparameter** details (IPs)

- **Research question**:
  - can an adversary still **infer architecture** and optimization **hyperparameters** ?



Experimental Setup

- MNIST black box classifiers

- Three model (hyper)parameter types:
  - (1) architecture
  - (2) optimization
  - (3) training data

- Ask adversary multiple-choice questions:
  - e.g.: “Which of the following activation functions does this black box model use? [ReLU, PReLU, ELU, Tanh]”

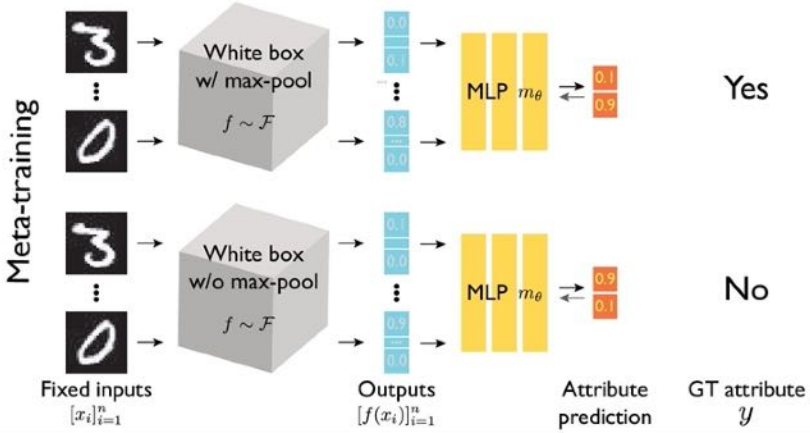
	Code	Attribute	Values
Architecture	act	Activation	ReLU, PReLU, ELU, Tanh
	drop	Dropout	Yes, No
	pool	Max pooling	Yes, No
	ks	Conv ker. size	3, 5
	#conv	#Conv layers	2, 3, 4
	#fc	#FC layers	2, 3, 4
	#par	*#Parameters*	$2^{14}, \cdots, 2^{21}$
	ens	Ensemble	Yes, No
Opt.	alg	Algorithm	SGD, ADAM, RMSprop
	bs	Batch size	64, 128, 256
Data	split	Data split	All_0, $\text{Half}_{0/1}$, $\text{Quarter}_{0/1/2/3}$
	size	*Data size*	All, Half, Quarter

Method Overview: `kennen`

- “Kennen”: to know (German) or to dig out (Korean)

- Hypothesis:
  - **model outputs** contain **fingerprints** of internal (hyper)parameters

- Approach:
  - train 5,000 diverse white box MNIST classifiers covering all hyperparameters
  - learn to classify hyperparameters using sets of input / output pairs of the 5,000 white-box models
  - apply classifier to unseen black-box models to predict their hyperparameters.



Results

- Positive:
  - kennen-io achieves 80.1% acc (1,000 queries, score outputs, 5k models).
  - for architecture and optimization (hyper)parameters

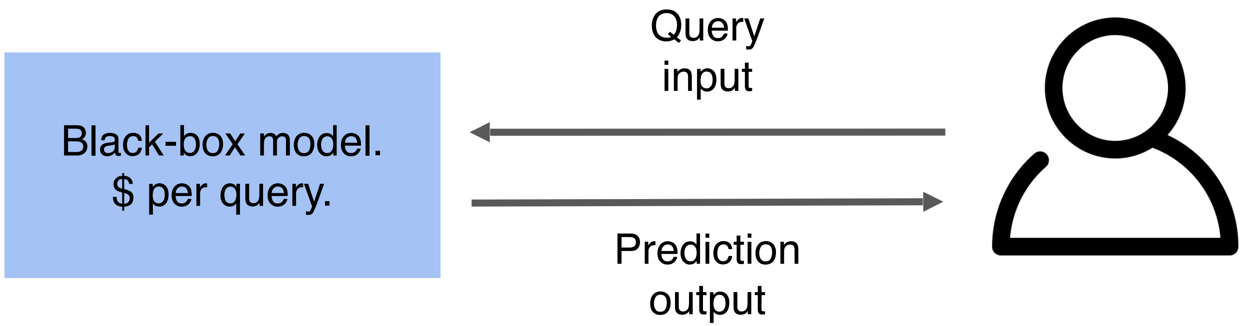
		architecture								optim		data		
Method	Output	act	drop	pool	ks	#conv	#fc	#par	ens	alg	bs	size	split	avg
Chance	-	25.0	50.0	50.0	50.0	33.3	33.3	12.5	50.0	33.3	33.3	33.3	14.3	34.9
`kennen-o`	prob	80.6	94.6	94.9	84.6	67.1	77.3	41.7	54.0	71.8	50.4	73.8	90.0	73.4
`kennen-o`	ranking	63.7	93.8	90.8	80.0	63.0	73.7	44.1	**62.4**	65.3	47.0	66.2	86.6	69.7
`kennen-o`	bottom-1	48.6	80.0	73.6	64.0	48.9	63.1	28.7	52.8	53.6	41.9	45.9	51.4	54.4
`kennen-o`	top-1	31.2	56.9	58.8	49.9	38.9	33.7	19.6	50.0	36.1	35.3	33.3	30.7	39.5
`kennen-i`	top-1	43.5	77.0	94.8	88.5	54.5	41.0	32.3	46.5	45.7	37.0	42.6	29.3	52.7
`kennen-io`	score	**88.4**	**95.8**	**99.5**	**97.7**	**80.3**	**80.2**	**45.2**	60.2	**79.3**	**54.3**	**84.8**	**95.6**	**80.1**

- Negative:
  - very costly (5k models)
  - scalability beyond MNIST?

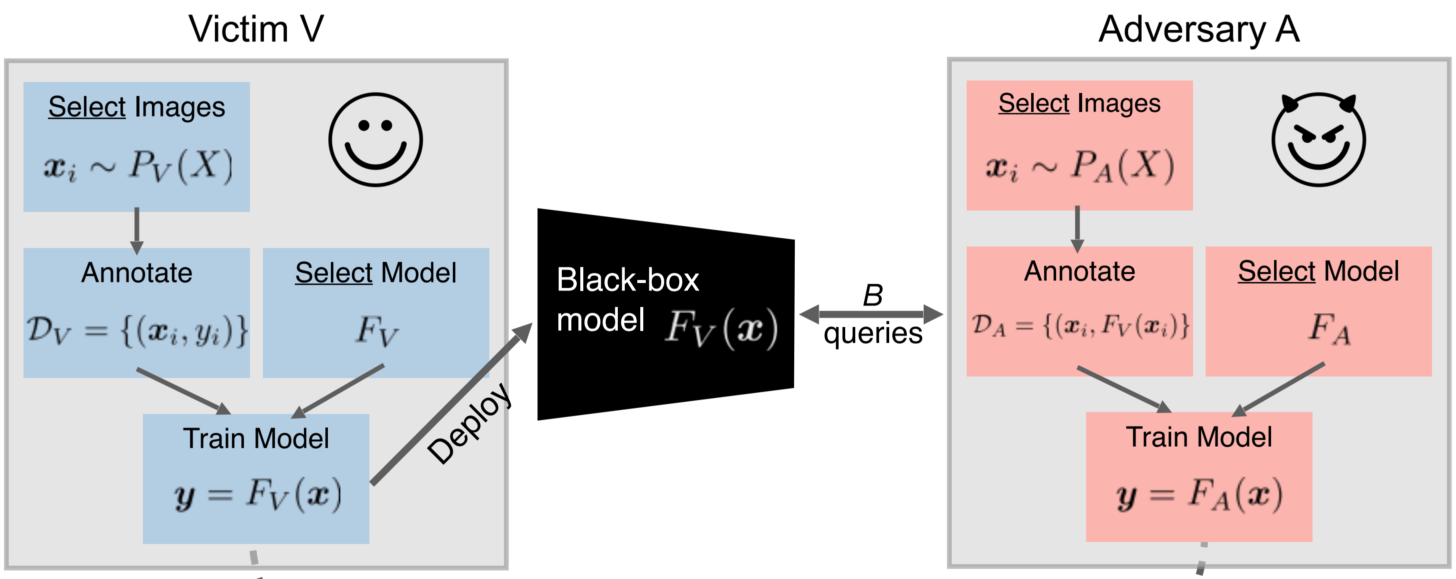
Providing ML Models is a Business Model

- **Input in, prediction out**. Ask $ per query.
  - ML models are **black boxes** !
  - not shared: **architecture**, **parameters**, **hyperparameter** details (IPs)

- **Research question**:
  - can an adversary **steal the functionality of the model** ?



Functionality Stealing: Knock-Off Nets (CVPR'19)



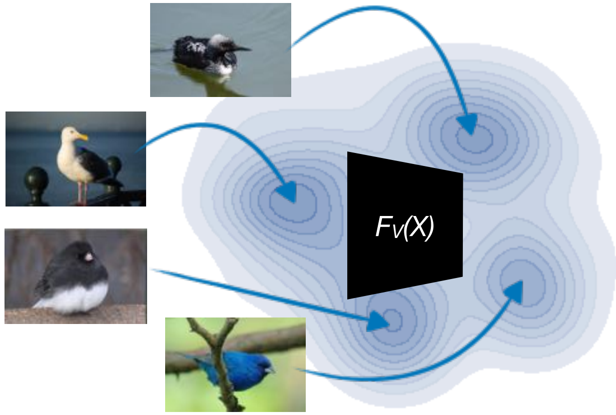
Can A "steal" F_V :

1. **when P_V is unknown?**
2. **when F_V architecture unknown?**
3. **using few queries B?**

Resembles Model Distillation … but under weaker assumptions

Query Set Selection: Challenge

$P_V(X)$



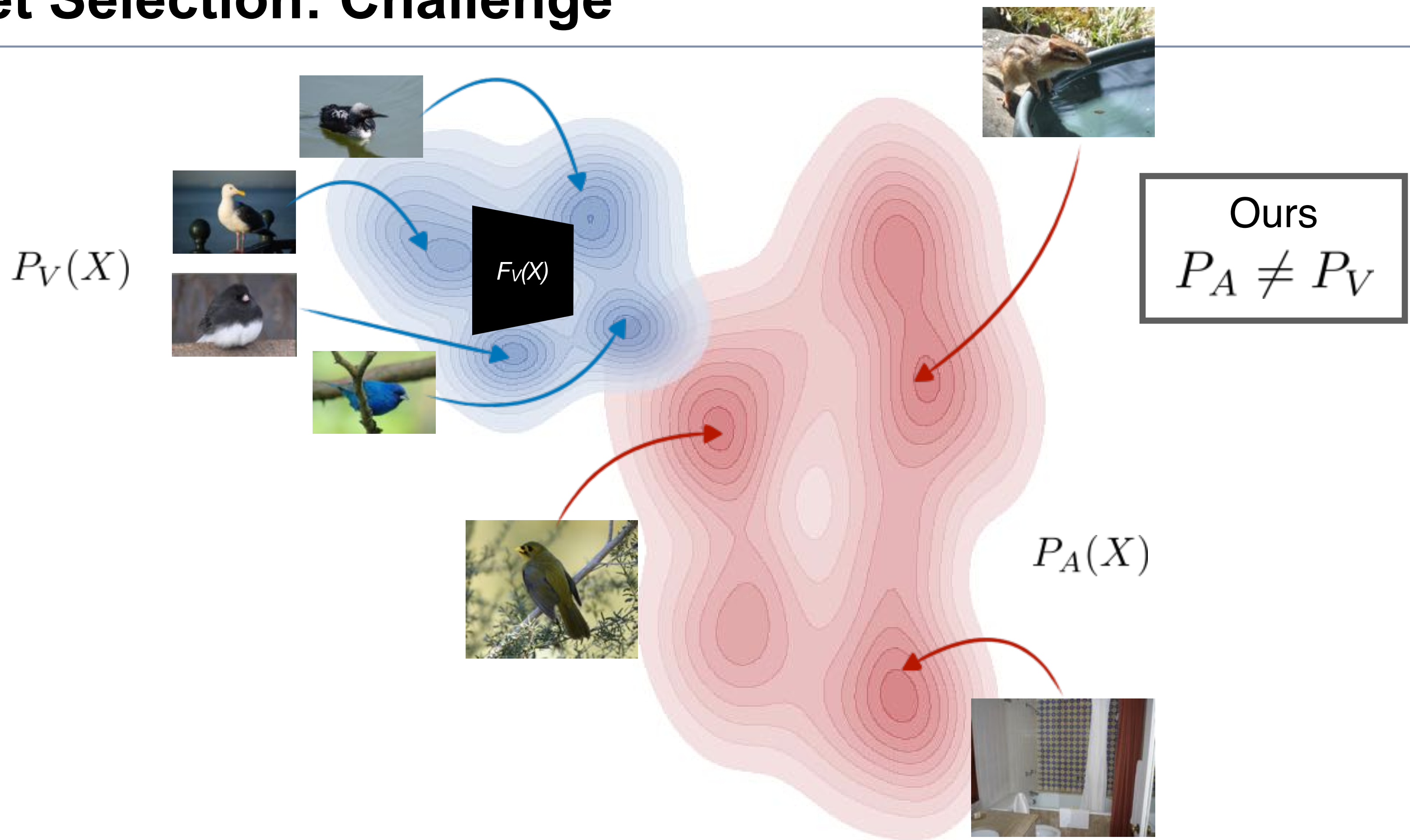
Active Learning

Distillation

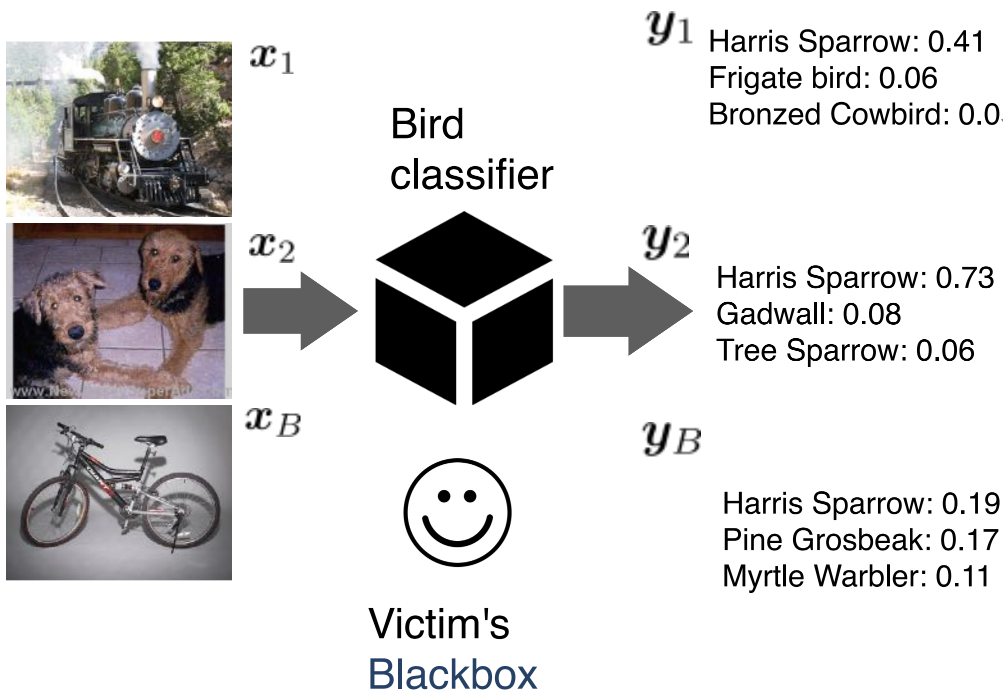
Student-Teacher

$$P_V = P_A$$

Query Set Selection: Challenge

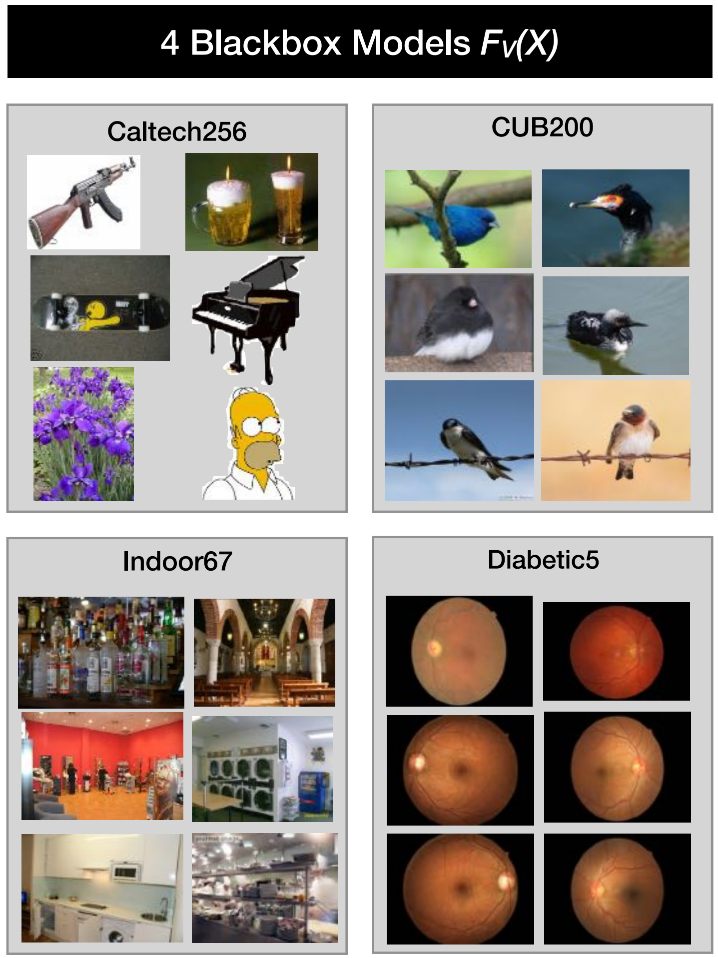


Functionality Stealing: Knock-Off Nets



Transfer Set Construction: $\boldsymbol{x}_i \overset{\pi}{\sim} P_A(X)$

- Simple method: π = random
  - sample images randomly (without replacement)
  - prone to querying irrelevant images



Can we Learn with π = Random? Yes!

	P_A	random			
		Caltech256	CUBS200	Indoor67	Diabetic5
	$P_V(F_V)$	78.8 (1×)	76.5 (1×)	74.9 (1×)	58.1 (1×)
	P_V (KD)	82.6 (1.05×)	70.3 (0.92×)	74.4 (0.99×)	54.3 (0.93×)
Closed	D^2	76.6 (0.97×)	68.3 (0.89×)	68.3 (0.91×)	48.9 (0.84×)
Open	ILSVRC	75.4 (0.96×)	68.0 (0.89×)	66.5 (0.89×)	47.7 (0.82×)
	OpenImg	73.6 (0.93×)	65.6 (0.86×)	69.9 (0.93×)	47.0 (0.81×)

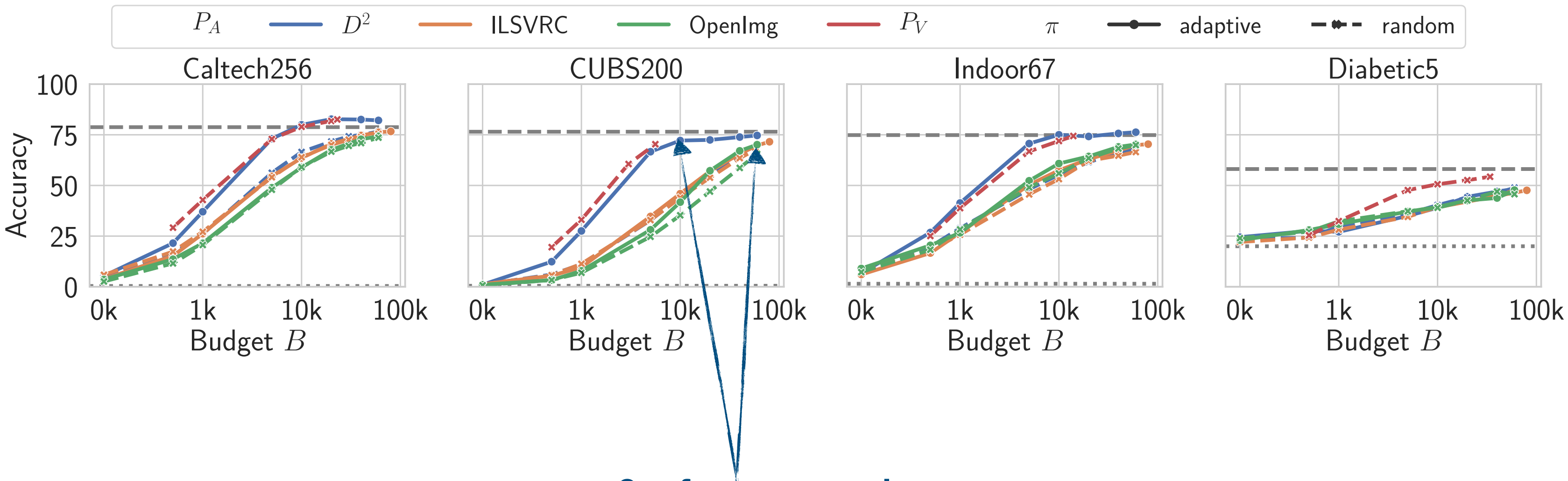
accuracy(victim blackbox)

accuracy(knockoff)

⇒ > 0.81× accuracy of blackbox recovered

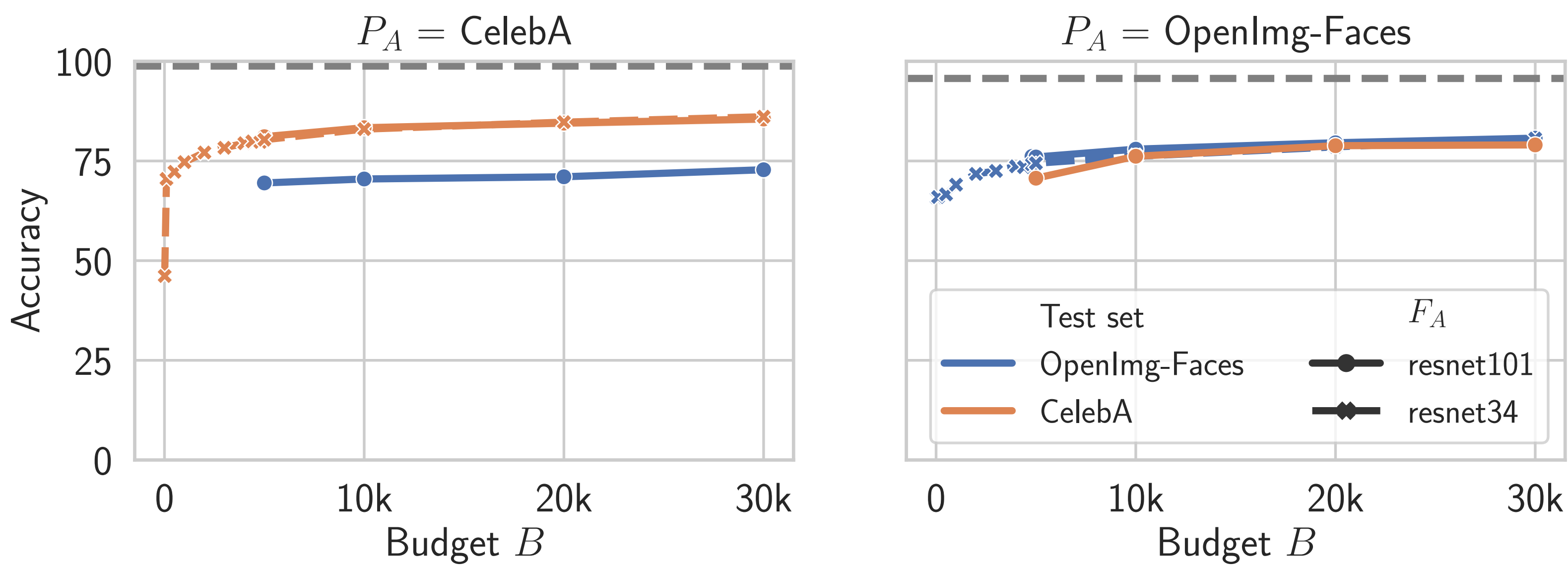
Can Make it Sample-Efficient? Yes!

		random				adaptive			
	P_A	Caltech256	CUBS200	Indoor67	Diabetic5	Caltech256	CUBS200	Indoor67	Diabetic5
	$P_V (F_V)$	78.8 (1×)	76.5 (1×)	74.9 (1×)	58.1 (1×)	-	-	-	-
	P_V (KD)	82.6 (1.05×)	70.3 (0.92×)	74.4 (0.99×)	54.3 (0.93×)	-	-	-	-
Closed	D^2	76.6 (0.97×)	68.3 (0.89×)	68.3 (0.91×)	48.9 (0.84×)	82.7 (1.05×)	74.7 (0.98×)	76.3 (1.02×)	48.3 (0.83×)
Open	ILSVRC	75.4 (0.96×)	68.0 (0.89×)	66.5 (0.89×)	47.7 (0.82×)	76.2 (0.97×)	69.7 (0.91×)	69.9 (0.93×)	44.6 (0.77×)
	OpenImg	73.6 (0.93×)	65.6 (0.86×)	69.9 (0.93×)	47.0 (0.81×)	74.2 (0.94×)	70.1 (0.92×)	70.2 (0.94×)	47.7 (0.82×)



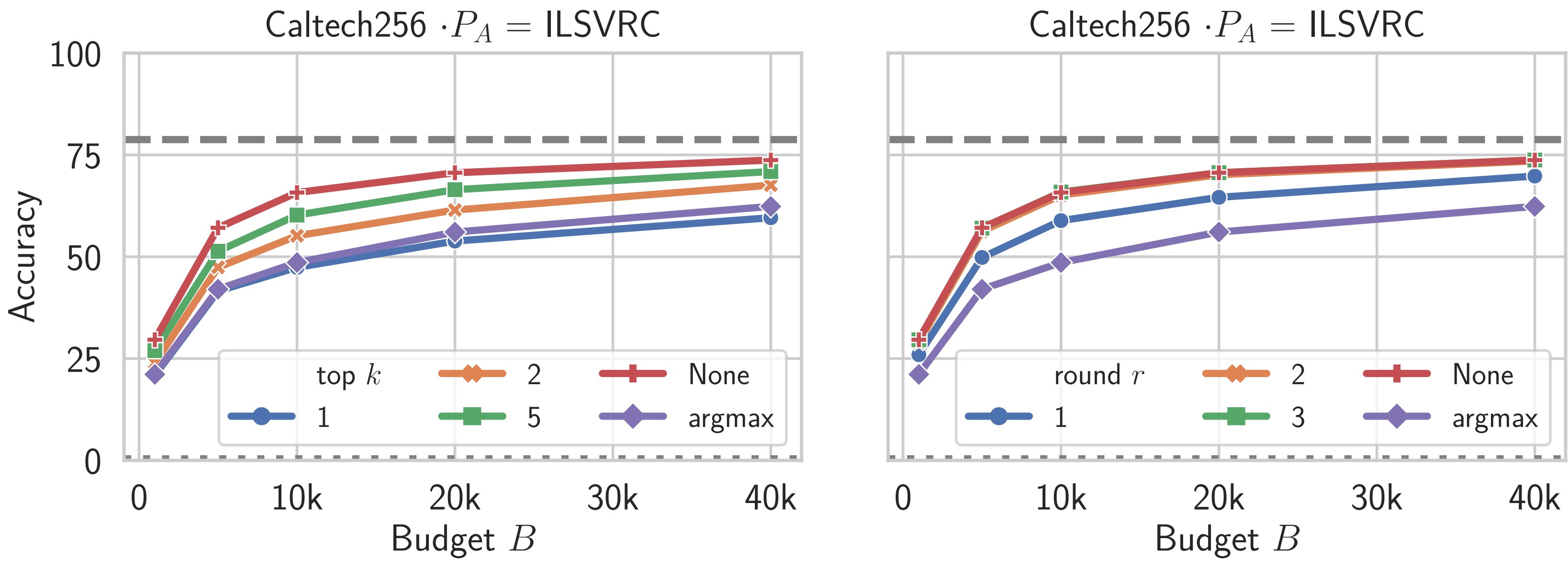
⇒ 6× fewer queries

Transfers to Real-World? Yes!



⇒ Also transfers to real-world API

Learning with Less Information? Yes!



⇒ Robust to various **passive** defense mechanisms:

e.g. argmax, top-k, rounding, …

Take Home Message - Stealing Deep Models…

- **Deep models** contain **intellectual property**
  - **model** and **learning parameters**
  - also **training** and **annotation** data

- Deploying deep models as a **black box** through an **API**
  - allows to **estimate model** and **learn parameters** (far beyond chance level)
  - allows to **steal** the **model's functionality reliably**
    - a few 1,000 queries are sufficient (or a few $)

  - unfortunately **difficult to defend — open research question**
    - passive defense: noising, top-k, argmax, rounding, … not particularly effective
    - active defense: "**prediction poisoning**"

max planck institut
informatik

SIC Saarland Informatics Campus

Challenges for Deep Learning in Computer Vision:

Interpretability, Robustness and Security

Bernt Schiele

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