

# Trustworthy and Explainable Artificial Intelligence

Anh Nguyen

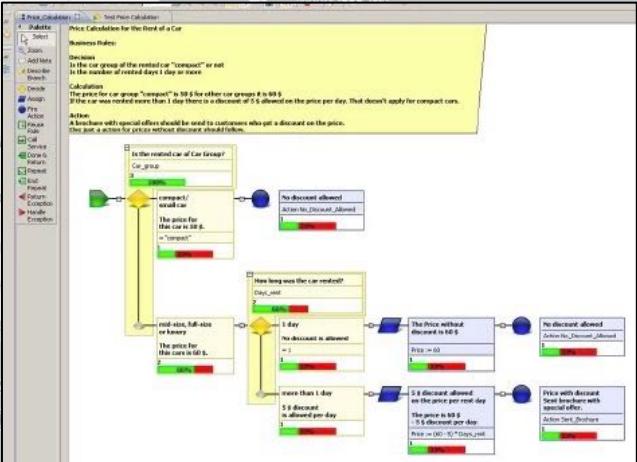
@anh\_ng8

<http://anhnguyen.me>



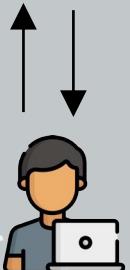


# Research Area: Trustworthy and Explainable Artificial Intelligence



Editor for Business Rules

Software



Business Executive

2009 Front-end programming

UBER AI Labs



# Research Area: Trustworthy and Explainable Artificial Intelligence

2014 M.S. in Human-Computer Interaction

## How An Intelligent Thimble Could Replace the Mouse In 3D Virtual Reality Worlds

Humans need better ways to interact with computers in three dimensions. Computer scientists think they've come up with one.

By Emerging Technology from the arXiv

The way in which humans interact with computers has been dominated by the mouse since it was invented in the 1960s by Doug Engelbart. A mouse uses a flat two-dimensional surface as a proxy for a computer screen. Any movements of the mouse over the surface are then translated into movements on the screen. These days, a mouse also has a number of buttons, and often scroll wheel, that allow interaction with on-screen objects.



GN TECH WORK LIFE NEWS IMPACT PODCASTS VIDEO RECOMMENDER INNOVATION FEST

07-09-14

FAST COMPANY

## Is A Smart Thimble The Computer Mouse Of The Future?

A new invention points a magic finger in the direction of the 3-D interfaces of tomorrow.

BY JOHN BROWNLEE 1 MINUTE READ

Although it has gotten us through the last 50 years of computer history, the computer mouse has one big, contemporary design problem: It only works well with 2-D interfaces. That's fine for a conventional desktop operating system, such as Windows or the Mac OS, but for the 3-D interfaces of the future, we're going to have to leave the mouse behind.

Anh Nguyen and Amy Banic think they know what will take its



Sensors = IMU + optical CMOS

- 01 paper at IEEE Virtual Reality 2015
- 01 patent
- Industry commercialization interests

# Research Area: Trustworthy and Explainable Artificial Intelligence

2017 Ph.D. in Artificial Intelligence



Jeff Clune



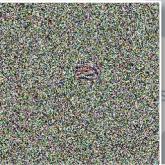
# Research Area: Trustworthy and Explainable Artificial Intelligence

2017 – present Assistant Professor



# Research Area: Trustworthy and Explainable Artificial Intelligence

## Trustworthy AI



School bus



Q1 Does marijuana cause cancer?

Q2'' lung can give marijuana smoking How you cancer?

(c) Prediction: "duplicate" 0.99

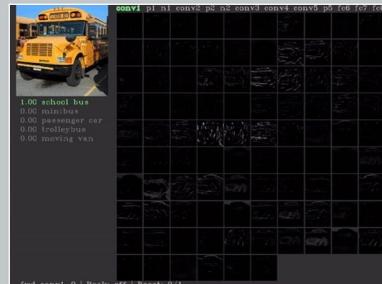
Q1 Does marijuana cause cancer?

Q1' Does cancer cause marijuana?

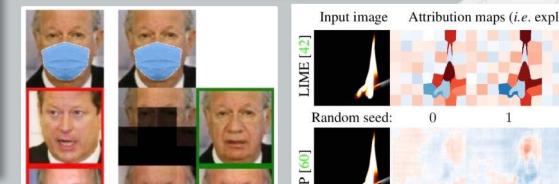
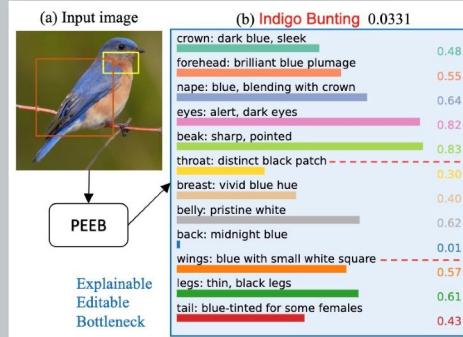
(d) Prediction: "duplicate" 0.77

2017 – present Assistant Professor

## Explainable AI



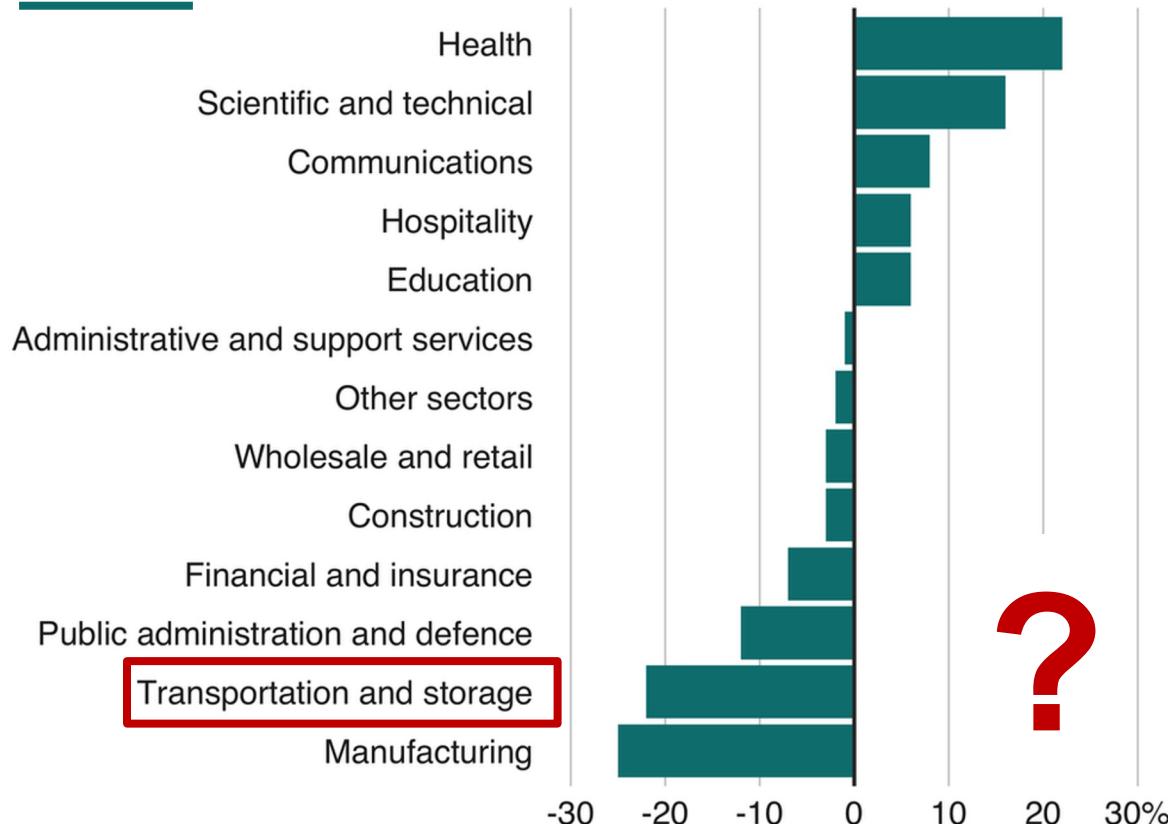
junco



AI is everywhere

# How AI could change the job market

Estimated net job creation by industry sector, 2017-2037



**2019** Elon Musk

"Feature complete for full self-driving this year"



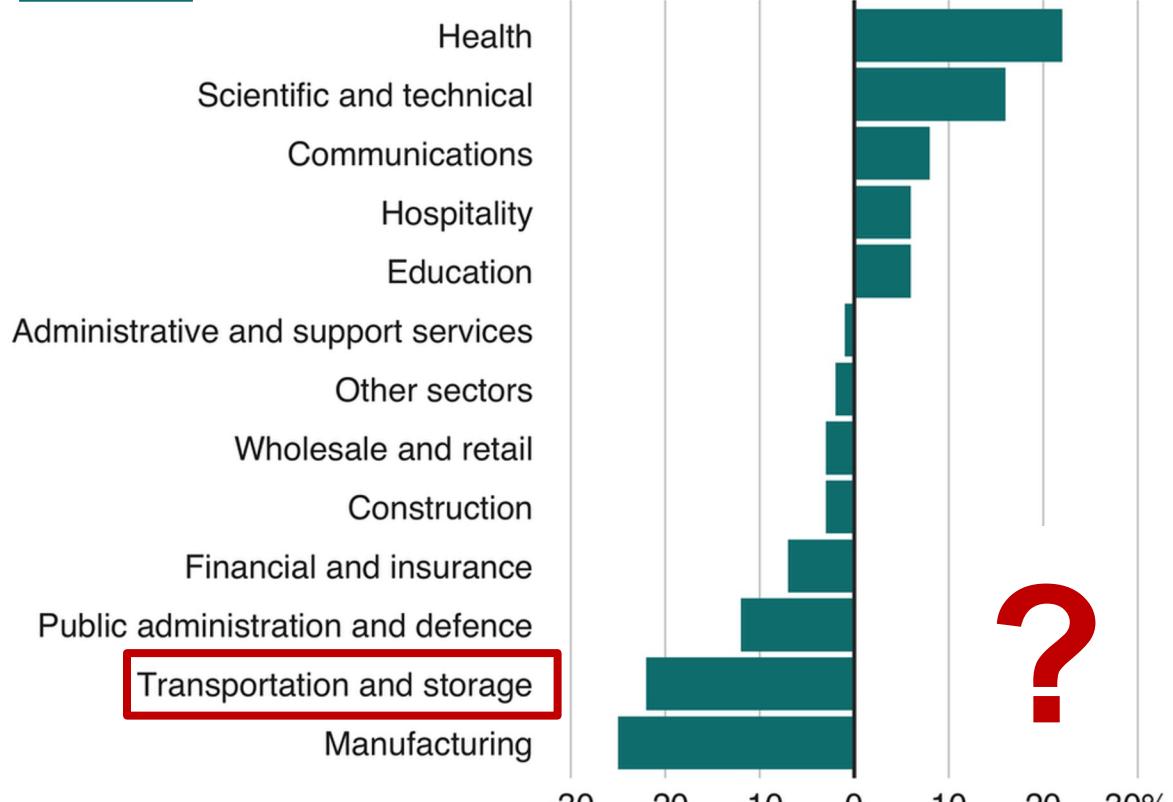
# How AI could change the job market

<https://twitter.com/giacaglia/status/1414605317841702914>

Estimated net job creation by industry sector, 2017-2037

2021

Tesla: the moon is a yellow traffic light



?



# 'Driverless' Tesla crash in Texas wasn't actually driverless, NTSB says

22

<https://twitter.com/giacaglia/status/1414605317841702914>

Two men were killed in what initially was thought to have been another example of Tesla customers abusing Autopilot

By Andrew J. Hawkins | @andyjayhawk | Oct 21, 2021, 3:16pm EDT

f t SHARE



Matt Dougherty/KHOU

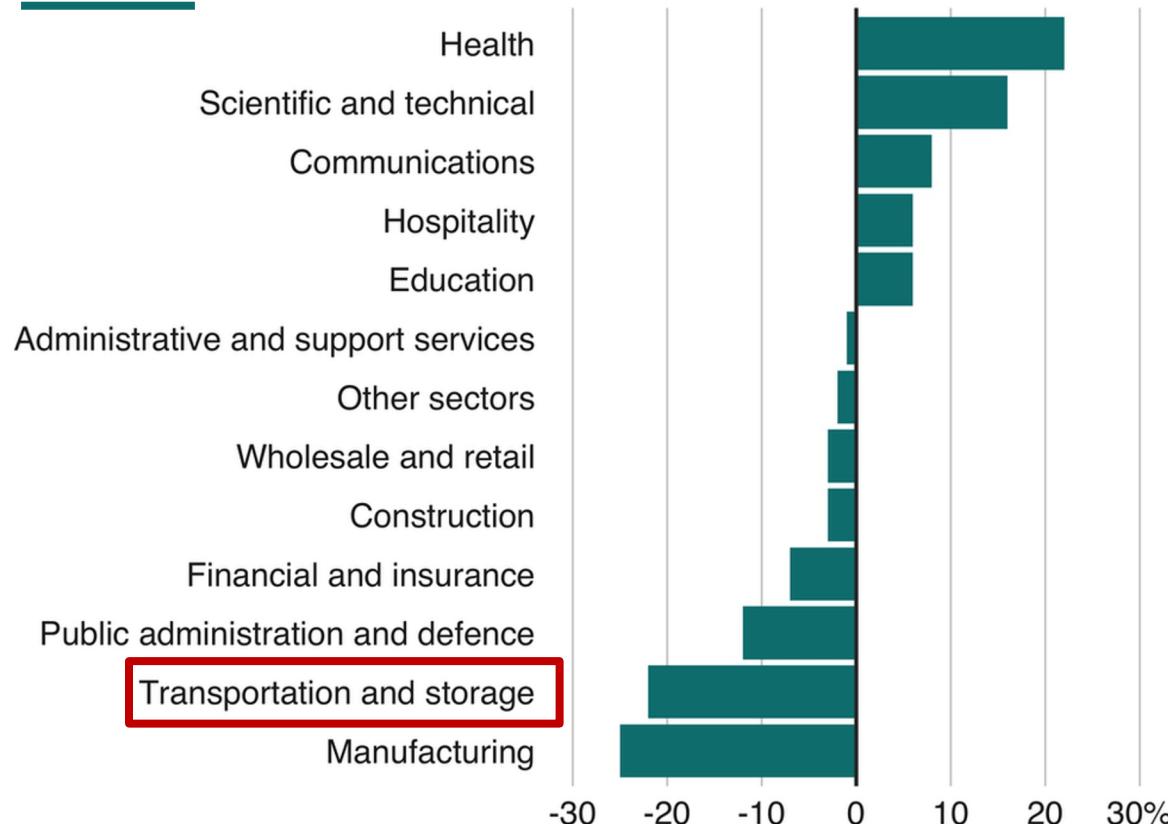
17

As of 07/2021, Tesla FSD 9.0:

1. Thinks the moon is a yellow traffic light and keeps slowing down
  2. Doesn't recognize planters in the street and almost hits them
  3. Doesn't recognize the monorails in the middle of the street and drives towards it!
  4. Goes through a bus lane instead of staying in the right lane
  5. Doesn't recognize a one-way street and the one-way sign in the street, and it drives towards the wrong way
  6. Can't decide which lane to use when turning a right in the next road. It keeps changing lanes
  7. Turns left but it goes to a lane that is not supposed to
- ...

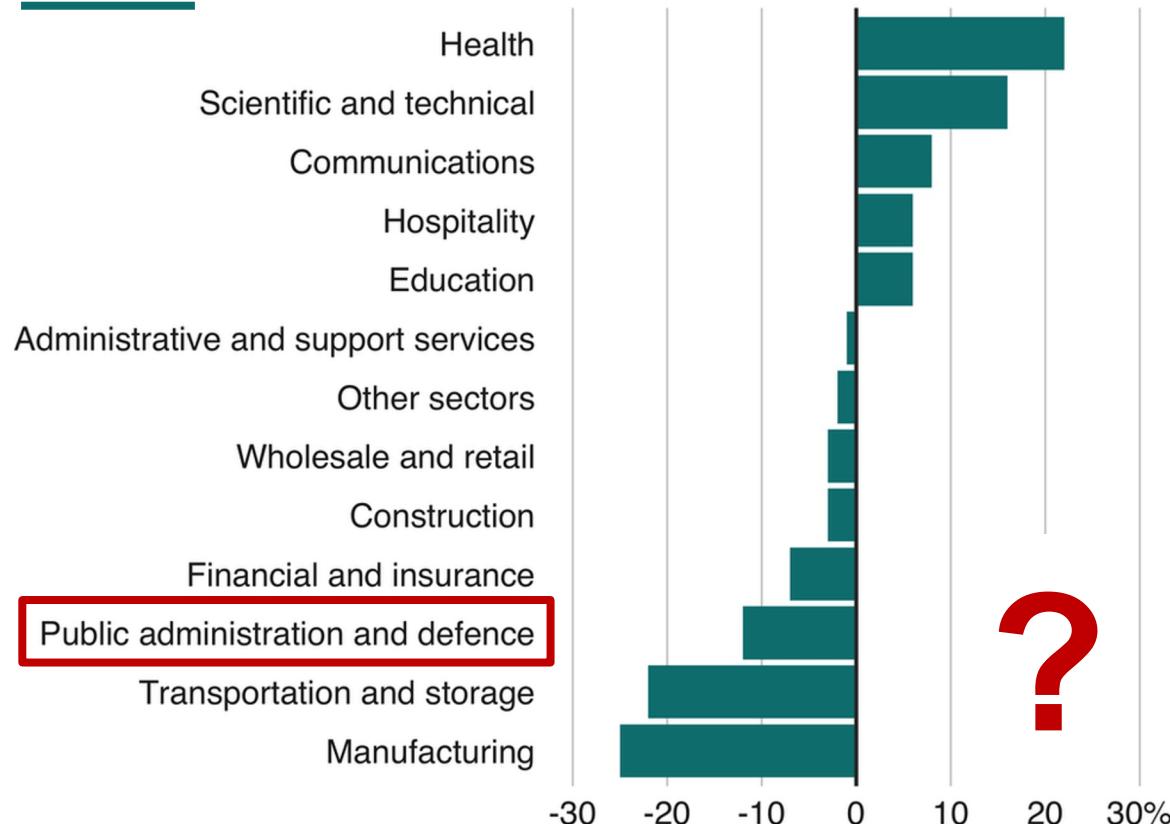
# How AI could change the job market

Estimated net job creation by industry sector, 2017-2037



# How AI could change the job market

Estimated net job creation by industry sector, 2017-2037



VERNON PRATER  
LOW RISK **3**

BRISHA BORDEN  
HIGH RISK **8**

TESLA / TECH

## Tesla sued by Texas cops after a Model X on Autopilot slammed into five officers

By Andrew J. Havens  
Sep 28, 2021, 10:30 AM

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

<https://www.theverge.com/2021/9/28/22698388/tesla-texas-lawsuit-cops-autopilot-crash-injury>

trivago

Search USA

Price / night \$0 - \$600+

Property type All Hotel House

Stays found: 1000+ Booking sites seen

Tru By Hilton Mobile

Hotel Mobile

\$131 View Details

# What does the GDPR and the "right to explanation" mean for AI?

Security teams increasingly rely on machine learning and artificial intelligence assets. Will a requirement to explain how they make decisions make them less

Query image

high

14



## Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match

A New Jersey man was accused of shoplifting and trying to hit an officer with a car. He is the third known Black man to be wrongfully arrested based on face recognition.



## *Wrongfully Accused by an Algorithm*

In what may be the first known case of its kind, a faulty facial

# The new lawsuit that shows facial recognition is officially a civil rights issue

Robert Williams, who was wrongfully arrested because of a faulty facial recognition match, is asking for the technology to be banned.

by Tate Ryan-Mosley

April 14, 2021

On January 9, 2020, Detroit police drove to the suburb of Farmington Hill and arrested Robert Williams in his driveway while his wife and young daughters looked on. Williams, a Black man, was accused of stealing watches from Shinola, a luxury store. He was held overnight in jail.

## Amazon's Face Recognition Falsey Members of Congress With Mugshots



By Jacob Snow, Technology & Civil Liberties Attorney, ACLU of M

JULY 26, 2018 | 8:00 AM

TAGS: Face Recognition Technology, Surveillance Technologies, Privacy



Amazon's face surveillance technology is the target of growing opposition nationwide, and today, there are 28 more causes for concern. In a test the ACLU recently conducted of the facial recognition tool, called "Rekognition," the software incorrectly matched 28



# Long-term Research Goals

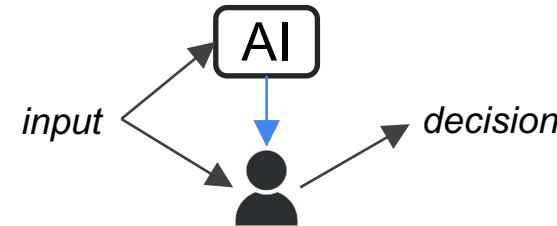
## Trustworthy AI

1. Build AIs that are accurate in edge cases  
(and common cases)

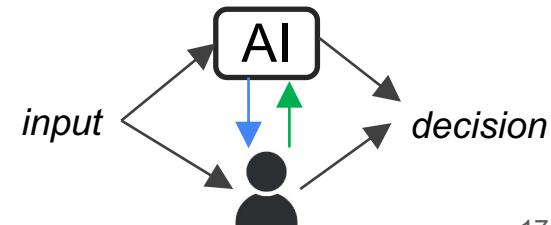


## Explainable AI

2. Build AIs that maximize human-AI team accuracy



3. Build AIs that humans can debug and edit  
(AI's decision-making process)



# Long-term Research Goals

## Trustworthy AI

1. Build ~~Test~~ AIs in edge cases  
(and common cases)

## Explainable AI

2. Build AIs that maximize human-AI team accuracy

3. Build AIs that humans can debug and edit  
(their decision-making process)

# Part 1

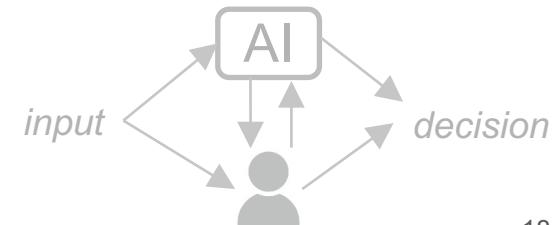
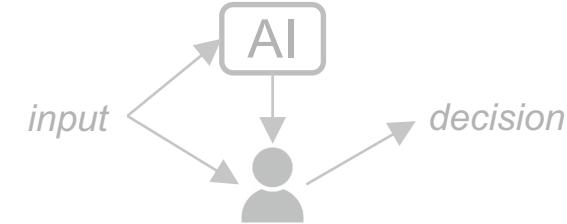


Image classifier

AI

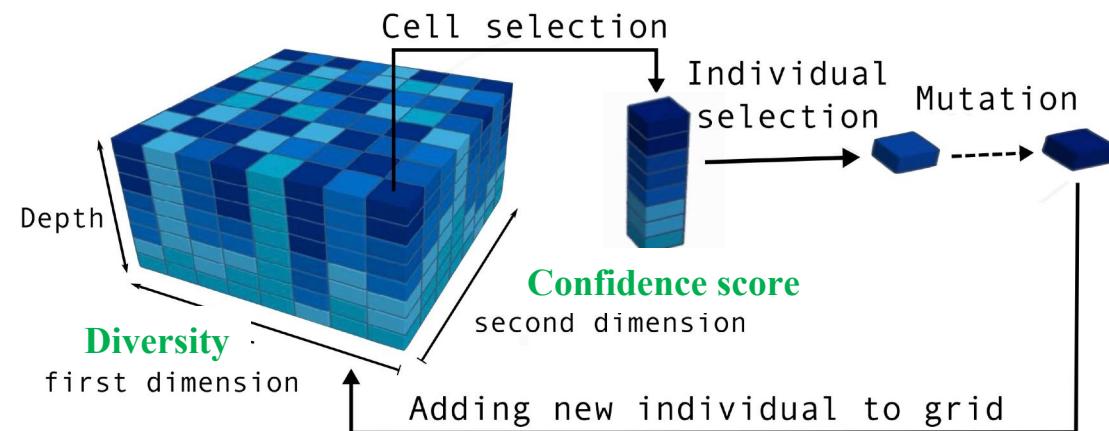
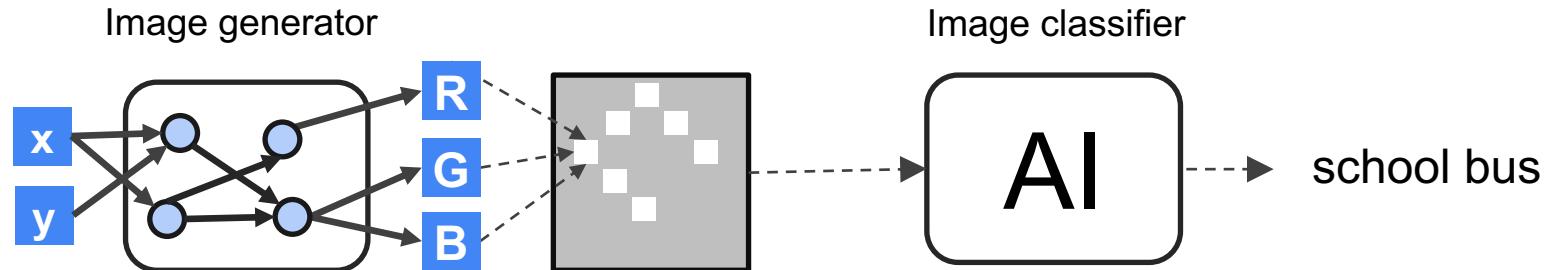
100%  
school bus?

School bus

training set



# A naïve attempt to draw a school bus that AI wants to see



Evolutionary Algorithm (MAP-Elites) evolves images that **maximize AI confidence scores + image diversity**

Image generator

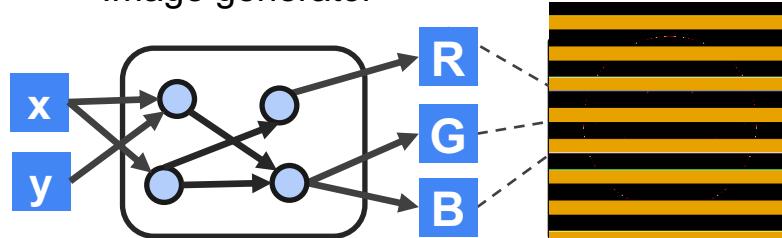
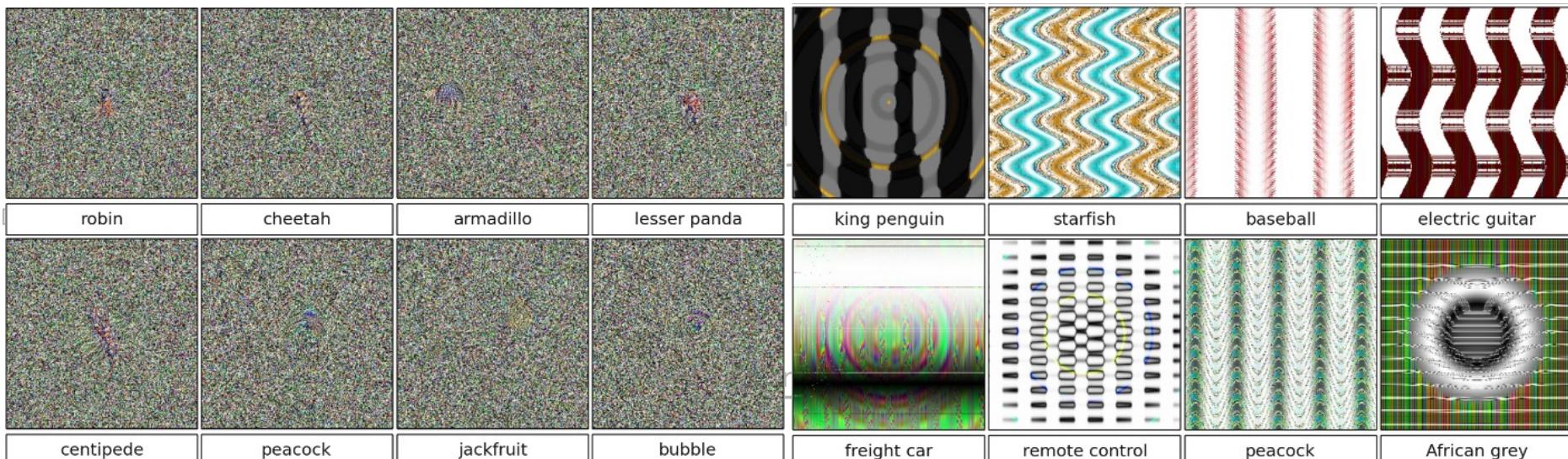


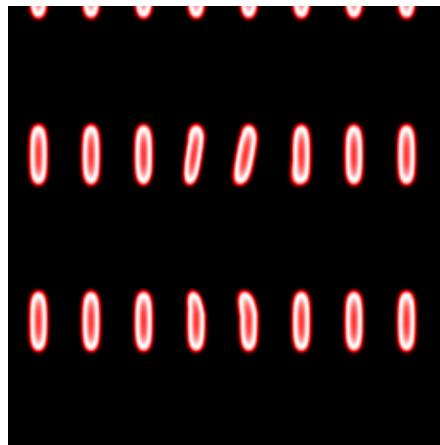
Image classifier

**99%**  
school bus

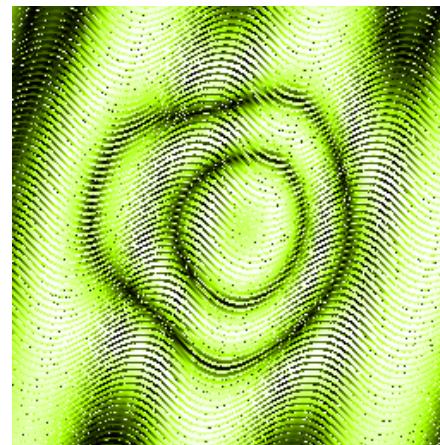


$\geq 96\%$  confidence

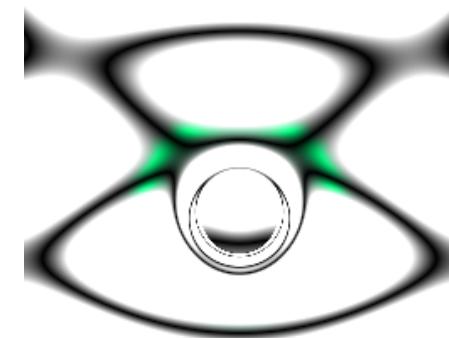
# How robust are fooling images?



digital clock



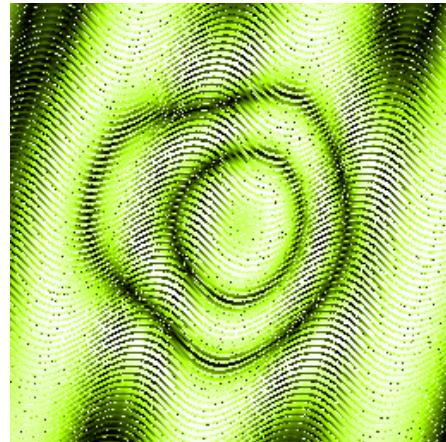
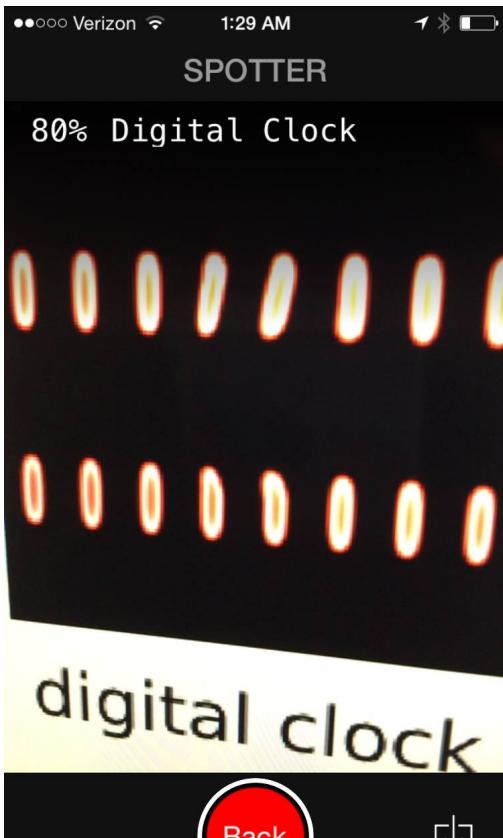
green snake



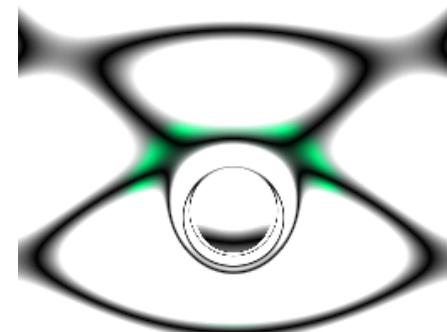
stethoscope

# How robust are fooling images?

Dileep George



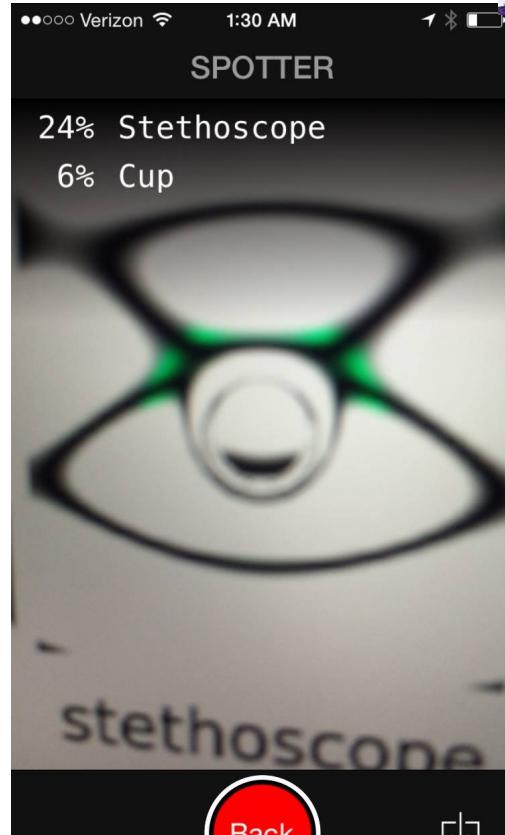
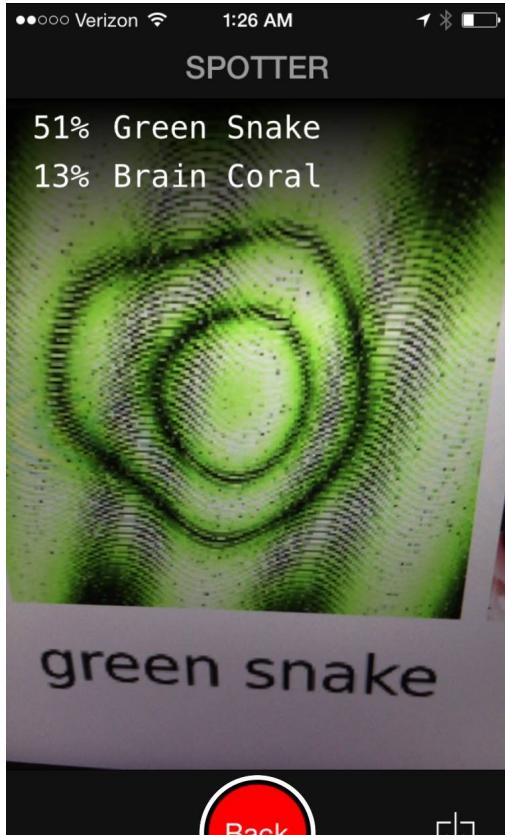
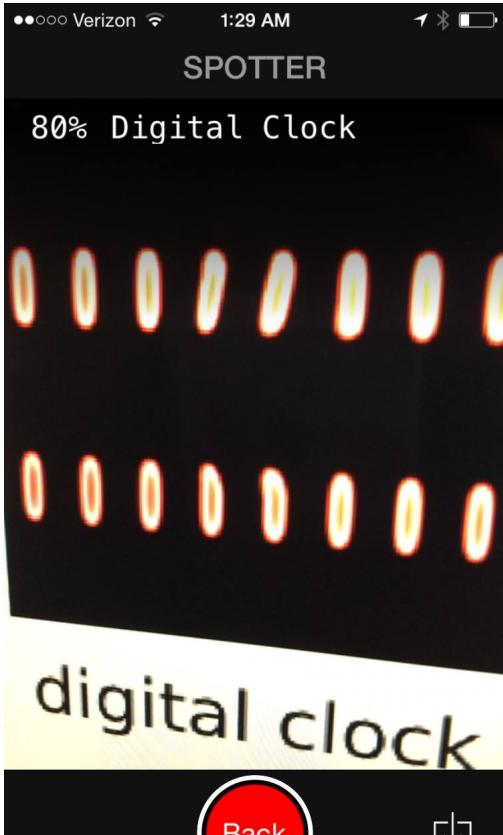
green snake



stethoscope

# How robust are fooling images?

Dileep George





99%  
school bus

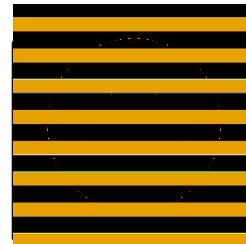
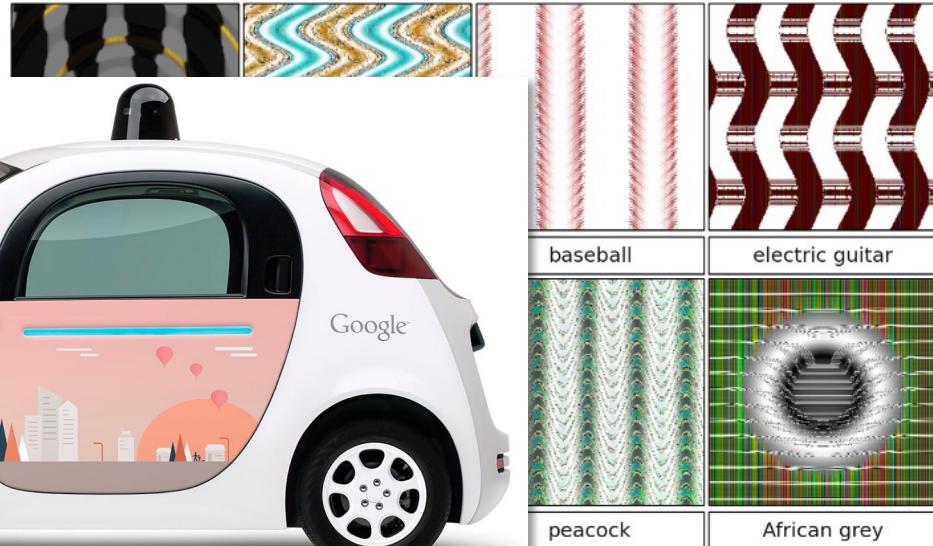


Image classifier

AI

99%  
school bus





- OpenCV **Top Paper Award** at CVPR 2015
- Altmetric **63<sup>rd</sup> most influential paper** worldwide in 2015

### Deep Neural Networks are Easily Fooled:

#### High Confidence Predictions for Unrecognizable Images

Anh Nguyen  
University of Wyoming  
anguyen8@uwyo.edu

Jason Yosinski  
Cornell University  
yosinski@cs.cornell.edu

Jeff Clune  
University of Wyoming  
jeffclune@uwyo.edu

#### Abstract

Deep neural networks (DNNs) have recently been achieving state-of-the-art performance on a variety of pattern-recognition tasks, most notably visual classification problems. Given that DNNs are now able to classify objects in images with near-human-level performance, questions naturally arise as to what differences remain between computer and human vision. A recent study [30] revealed that changing an image (e.g. of a lion) in a way imperceptible to humans can cause a DNN to label the image as something else entirely (e.g. mislabeling a lion a library). Here we show a related result: it is easy to produce images that are completely unrecognizable to humans, but that state-of-the-art DNNs believe to be recognizable objects with 99.99% confidence (e.g. labeling with certainty that white noise static is a lion). Specifically, we take convolutional neural networks trained to perform well on either the ImageNet or MNIST datasets and then find images with evolutionary algorithms or gradient ascent that DNNs label with high confidence as belonging to each dataset class. It is possible to produce images totally unrecognizable to human eyes that DNNs believe with near certainty are familiar objects, which we call “fooling images” (more generally, fooling examples). Our results shed light on interesting differences between human vision and current DNNs, and raise questions about the use of DNNs in applications.

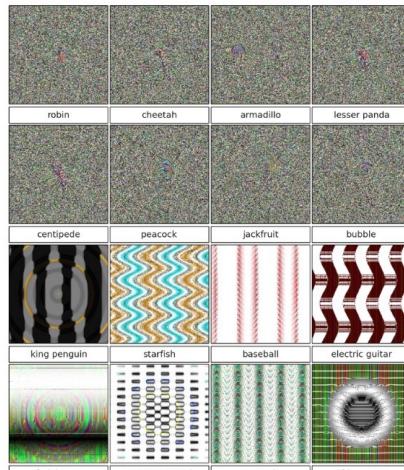


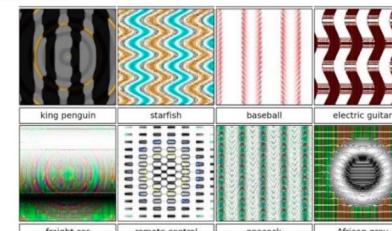
Figure 1. Evolved images that are unrecognizable to humans, but that state-of-the-art DNNs trained on ImageNet believe with  $\geq 99.6\%$  certainty to be a familiar object. This result highlights differences between how DNNs and humans recognize objects.

**MIT WIRELESS**  
**Technology Review**

**The Economist**

[Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images  
Nguyen, Yosinski, Clune, 2014]

>99.6%  
confidences



*The Atlantic*

**SCIENTIFIC AMERICAN MIND**  
IDEAS • BRAIN SCIENCE

# What about the text domain?

# NLU benchmark

- GLUE
- SuperGLUE
- Adversarial NLI
- MNLI + HANS

Human

Rank Name	Model	URL Score
1 DeBERTa Team - Microsoft	DeBERTa / TuringNLv4	 90.8
2 HFL iFLYTEK	MacALBERT + DKM	 90.7
+ 3 Alibaba DAMO NLP	StructBERT + TAPT	 90.6
+ 4 PING-AN Omni-Sinitic	ALBERT + DAAF + NAS	 90.6
5 ERNIE Team - Baidu	ERNIE	 90.4
6 T5 Team - Google	T5	 90.3
7 Microsoft D365 AI & MSR AI & GATECHMT-DNN-SMART		 89.9
+ 8 Huawei Noah's Ark Lab	NEZHA-Large	 89.8
+ 9 Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)	 89.7
+ 10 ELECTRA Team	ELECTRA-Large + Standard Tricks	 89.4
+ 11 Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)	 88.4
12 Junjie Yang	HIRE-RoBERTa	 88.3
13 Facebook AI	RoBERTa	 88.1
+ 14 Microsoft D365 AI & MSR AI	MT-DNN-ensemble	 87.6
15 GLUE Human Baselines	GLUE Human Baselines	 ICLR 2019 87.1

# NLU benchmark

- GLUE
- SuperGLUE
- Adversarial NLI
- MNLI + HANS

## BERT variants

Human

Rank Name	Model	URL Score
1 DeBERTa Team - Microsoft	DeBERTa / TuringNLv4	90.8
2 HFL iFLYTEK	MacALBERT + DKM	90.7
+ 3 Alibaba DAMO NLP	StructBERT + TAPT	90.6
+ 4 PING-AN Omni-Sinitic	ALBERT + DAAF + NAS	90.6
5 ERNIE Team - Baidu	ERNIE	90.4
6 T5 Team - Google	T5	90.3
7 Microsoft D365 AI & MSR AI & GATECHMT-DNN-SMART		89.9
+ 8 Huawei Noah's Ark Lab	NEZHA-Large	89.8
+ 9 Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)	89.7
+ 10 ELECTRA Team	ELECTRA-Large + Standard Tricks	89.4
+ 11 Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)	88.4
12 Junjie Yang	HIRE-RoBERTa	88.3
13 Facebook AI	RoBERTa	88.1
+ 14 Microsoft D365 AI & MSR AI	MT-DNN-ensemble	87.6
15 GLUE Human Baselines	GLUE Human Baselines	ICLR 2019

# NLU benchmark

- GLUE
- SuperGLUE
- Adversarial NLI
- MNLI + HANS

BERT variants  
outperformed  
Human

Rank Name	Model	URL	Score
1 DeBERTa Team - Microsoft	DeBERTa / TuringNLVRv4	ICLR 2021	90.8
2 HFL iFLYTEK	MacALBERT + DKM		90.7
+ 3 Alibaba DAMO NLP	StructBERT + TAPT	ICLR 2020	90.6
+ 4 PING-AN Omni-Sinitic	ALBERT + DAAF + NAS	*ICLR 2020	90.6
5 ERNIE Team - Baidu	ERNIE	ACL 2019	90.4
6 T5 Team - Google	T5	JMLR 2020	90.3
7 Microsoft D365 AI & MSR AI & GATECHMT-DNN-SMART			89.9
+ 8 Huawei Noah's Ark Lab	NEZHA-Large	ICLR 2020	89.8
+ 9 Zihang Dai	Funnel-Transformer (Ensemble B10-10)	NeurIPS 2020	89.7
+ 10 ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4
+ 11 Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)	ICLR 2020	88.4
12 Junjie Yang	HIRE-RoBERTa	*ICLR 2021	88.3
13 Facebook AI	RoBERTa	*ICLR 2020	88.1
+ 14 Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6
15 GLUE Human Baselines	GLUE Human Baselines	ICLR 2019	87.1

# GLUE

- Averaging normalized scores over 9 different tasks

We studied  
7 out of 9

Corpus	Train	Test	Task	Metrics	Domain
Single-Sentence Tasks					
CoLA	8.5k	<b>1k</b>	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks					
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	<b>391k</b>	paraphrase	acc./F1	social QA questions
Inference Tasks					
MNLI	393k	<b>20k</b>	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	<b>146</b>	coreference/NLI	acc.	fiction books

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

# GLUE

- Averaging normalized scores over 9 different tasks

Corpus	Train	Test	Task	Metrics	Domain
Single-Sentence Tasks					
MNLI	393k	<b>20k</b>	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	<b>146</b>	coreference/NLI	acc.	fiction books
Inference Tasks					

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

How can smoking marijuana give you lung cancer? → **AI**

Accuracy

*shuffling n-grams*

marijuana can smoking you how cancer give lung? → **AI**

Accuracy



How can smoking marijuana give you lung cancer? → **AI**

Accuracy

*shuffling n-grams*

**Marginal**  $\Delta$

marijuana can smoking you how cancer give lung? → **AI**

Accuracy



How can smoking marijuana give you lung cancer? → AI

Accuracy

*shuffling n-grams*

Marginal  $\Delta$

marijuana can smoking you how cancer give luna? → AI

Accuracy

2-noun swap 1-gram 2-gram 3-gram

---

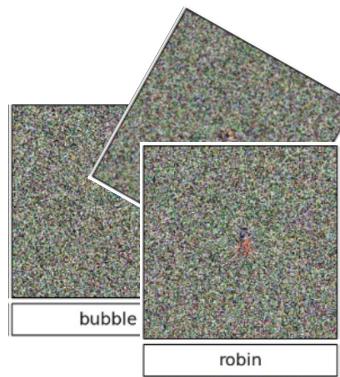
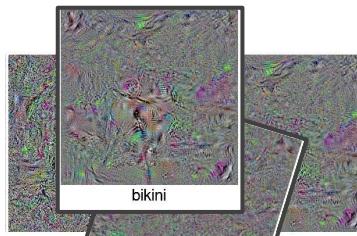
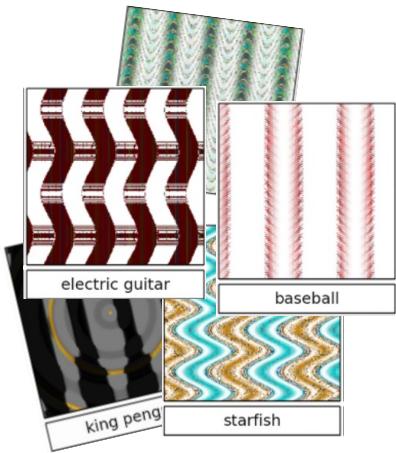
MRPC	100	50	96.51	83.89	87.1	89.38
------	-----	----	-------	-------	------	-------

---

SST-2	100	50	97.78	84.04	88.35	90.56
-------	-----	----	-------	-------	-------	-------

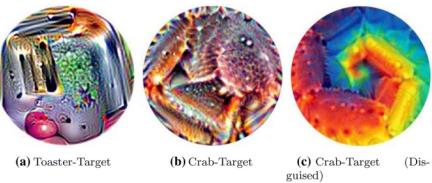
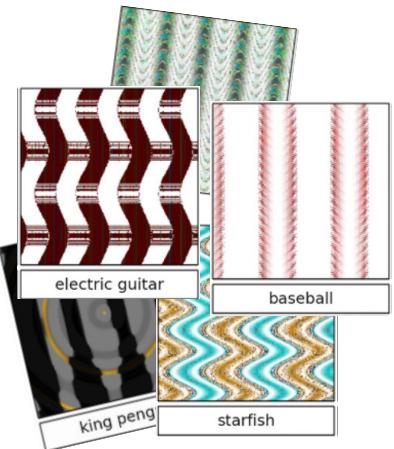
---

QNLI	100	50	94.31	89.42	93.85	95.32
------	-----	----	-------	-------	-------	-------

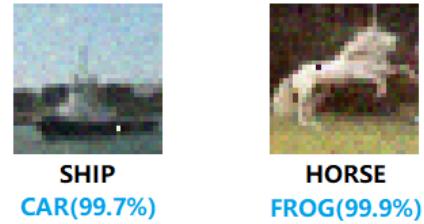
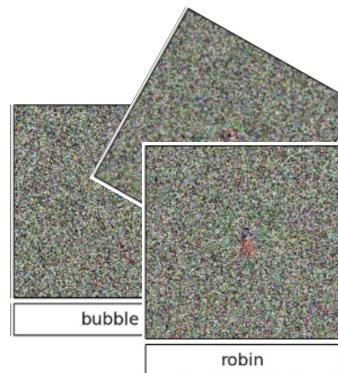
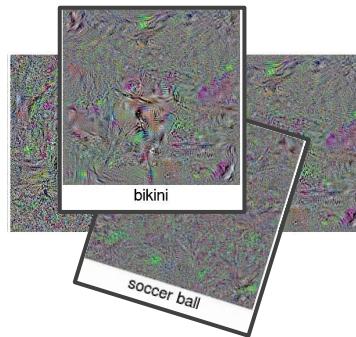




Universal perturbations



Adversarial Patch



One-pixel attack



Adversarial stop-sign

Can we fool vision models by a rotation?

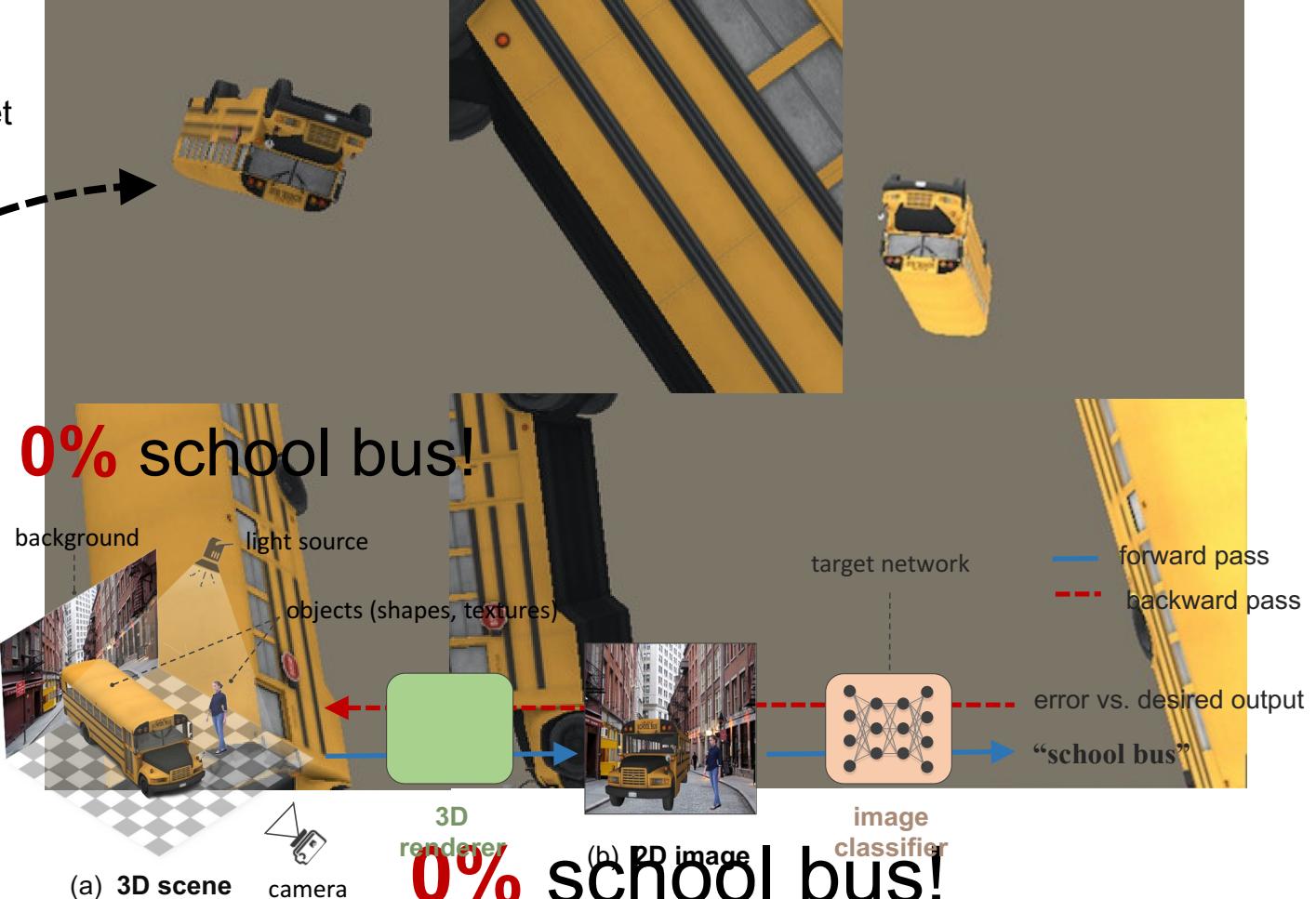
## Inception-v3

78% accuracy on ImageNet

i.i.d. test image



100%  
school bus



Inception-v3

78% accuracy on ImageNet

i.i.d. test image

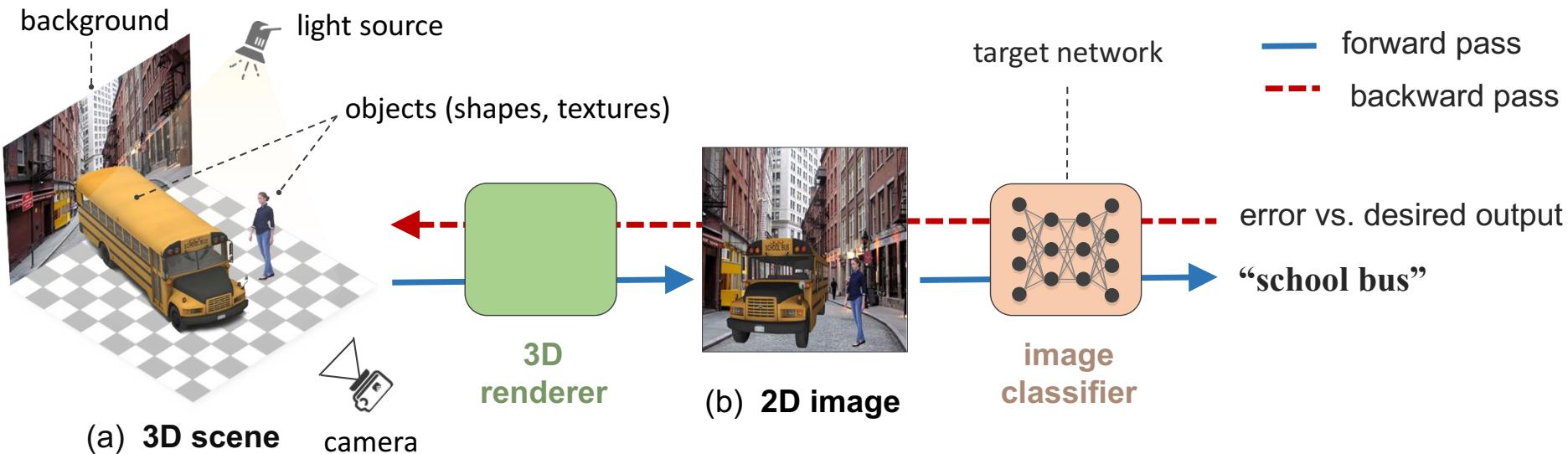


**100%**  
school bus



**100% confidence**

# Fine-grained control over stimuli changes



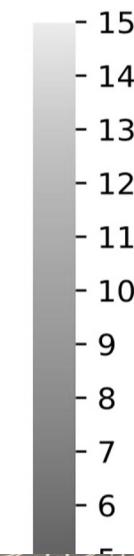
Alcorn et al. *Strike (with) a Pose*. CVPR 2019



Inception-v3

78% accuracy

-3.14  
-2.89  
-2.64



**school bus 1.0   garbage truck 0.99   punching bag 1.0   snowplow 0.92**

0.63  
0.88



umbrella 1.0

chain mail 0.83

chainlink fence 0.84

miniskirt 0.85

tripod 0.85

chain mail 0.88

quilt 0.88

chainlink fence 0.89

DNNs correctly label only 3% of the poses

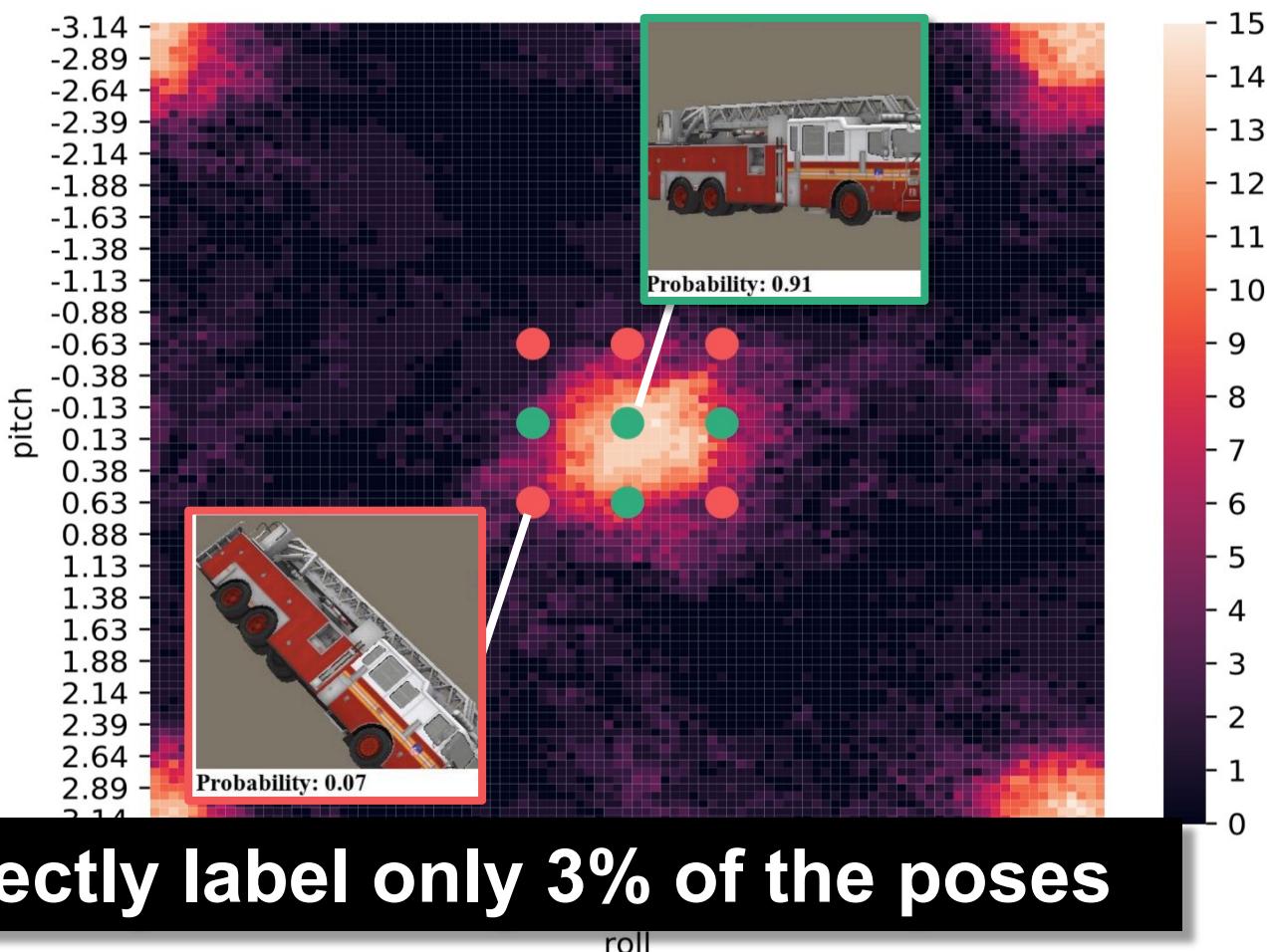
roll

Inception-v3

78% accuracy on ImageNet



100%

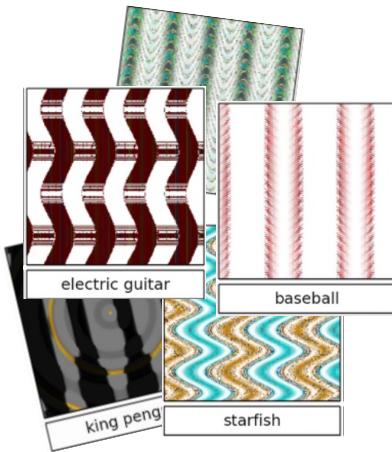


DNNs correctly label only 3% of the poses

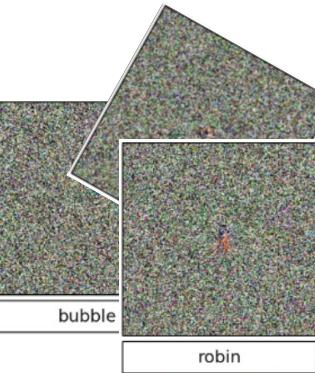
# Solution: Re-train AIs on misclassified examples?



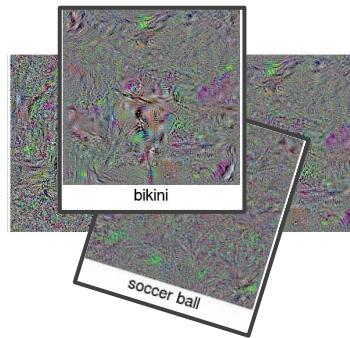
+



+



+

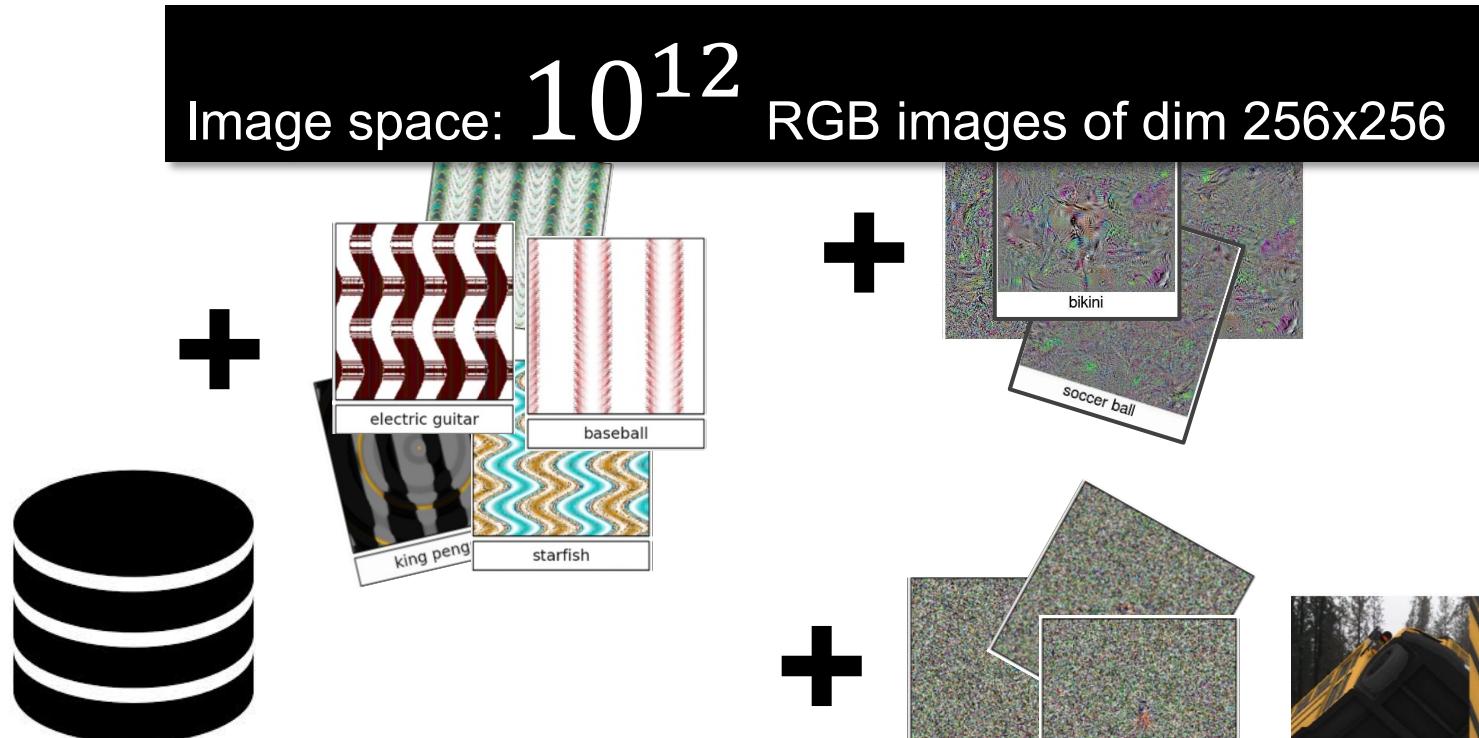


...



Training set

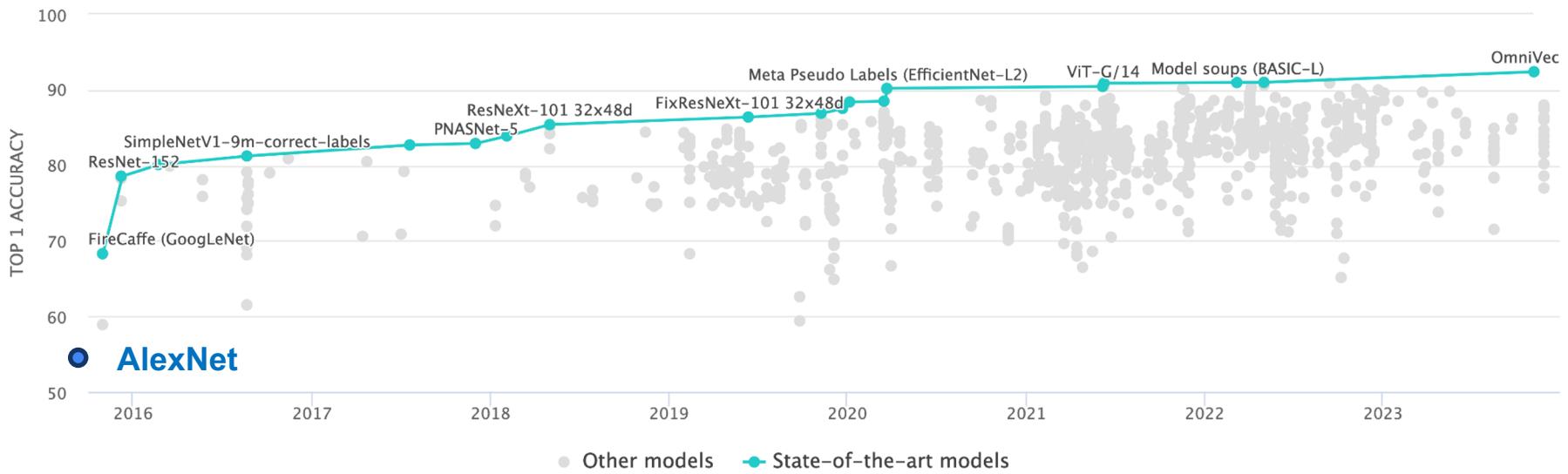
# Solution: Re-train AIs on misclassified examples? **Scalable?**



Training set

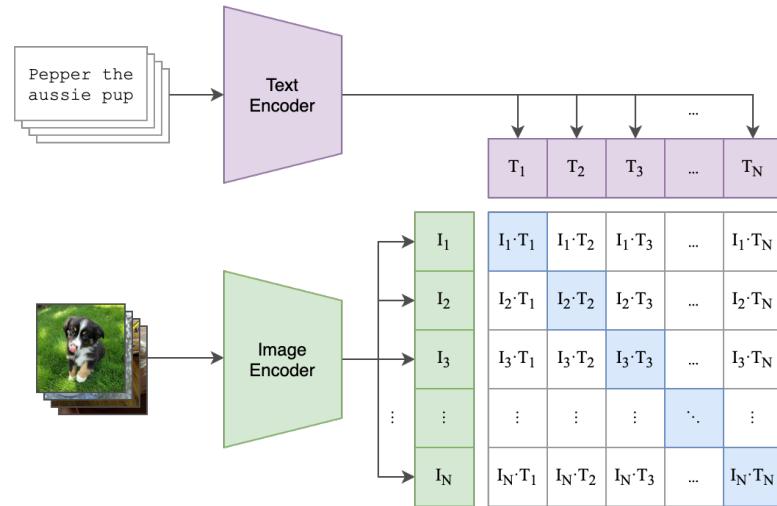
garbage truck 0.99 punching bag 1.0 snowplow 0

# ImageNet accuracy over time

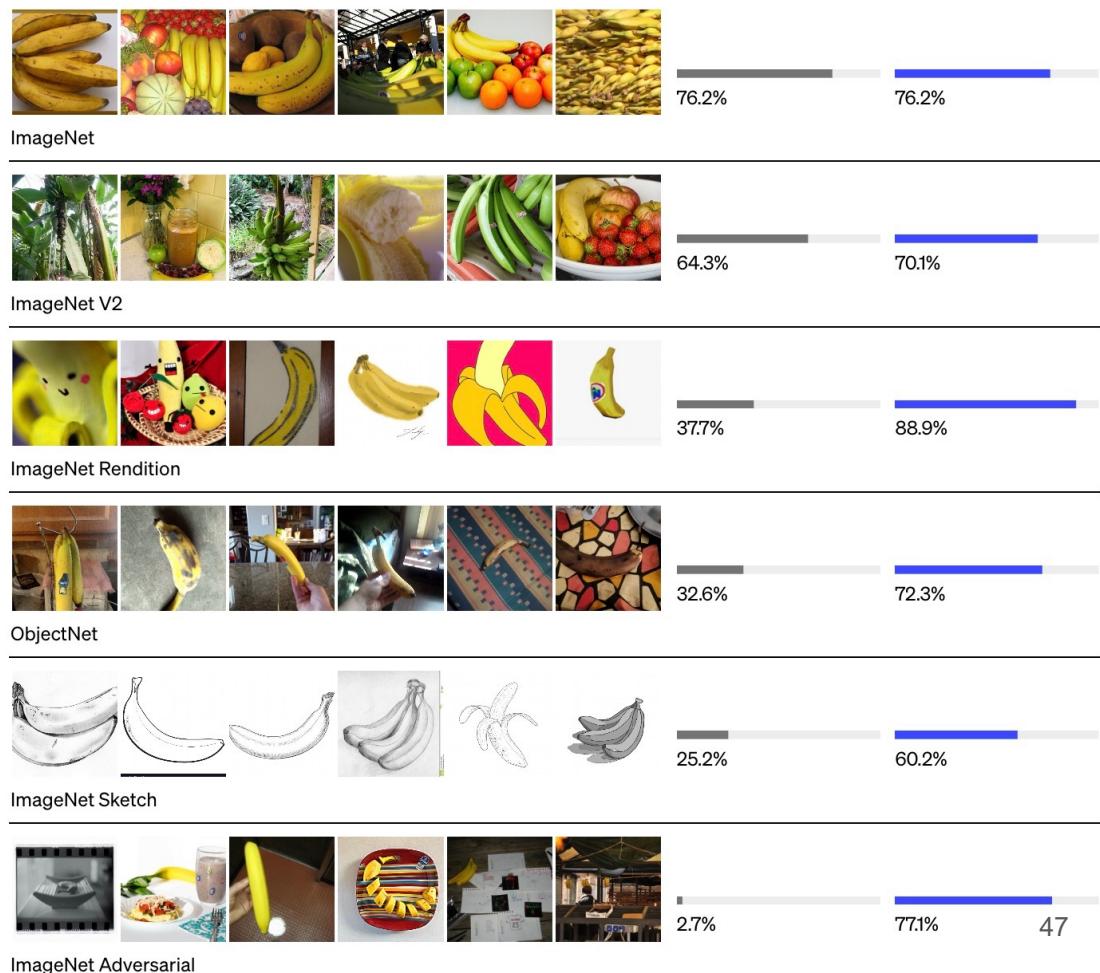


# OpenAI CLIP

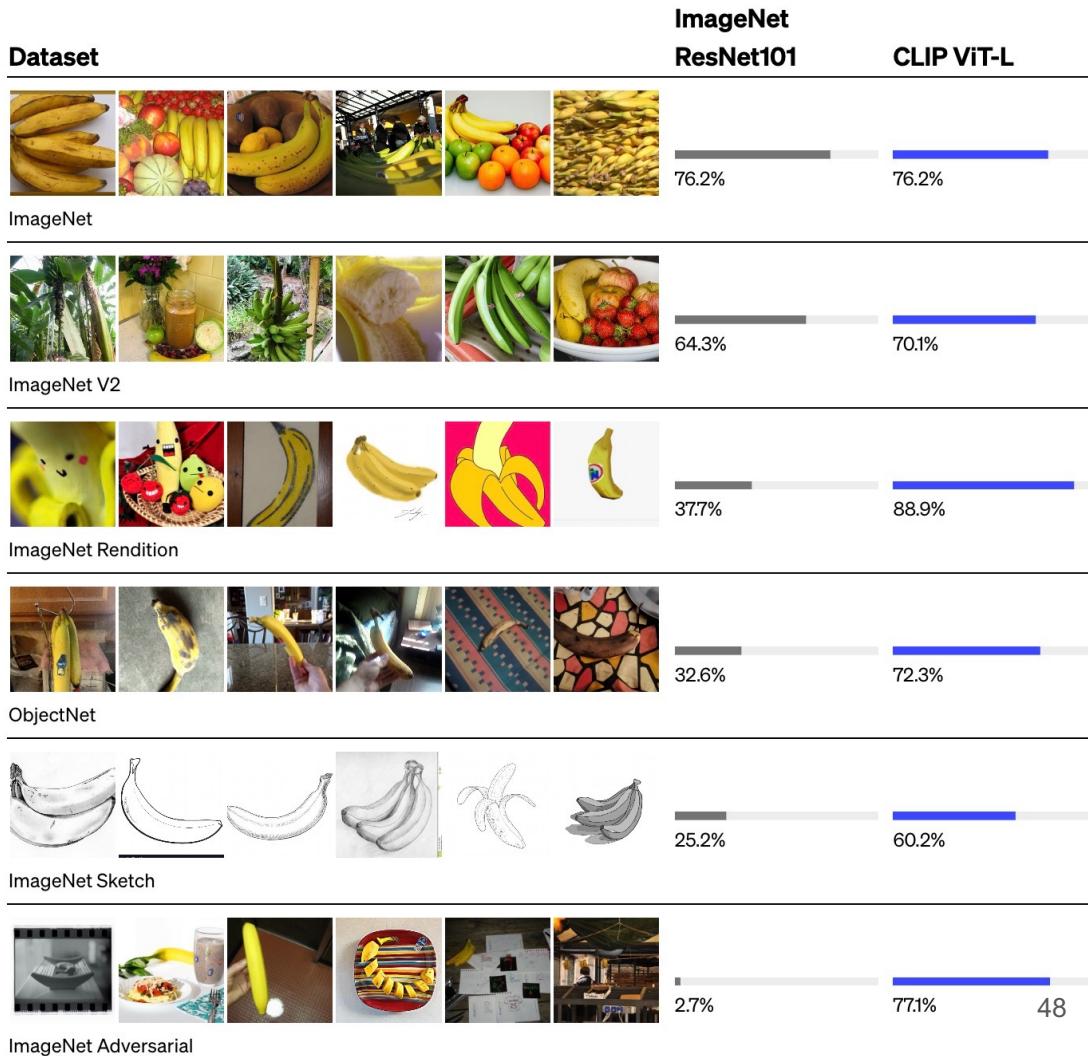
## (1) Contrastive pre-training



## Dataset



Q: The key challenge  
= zooming  
to the object?

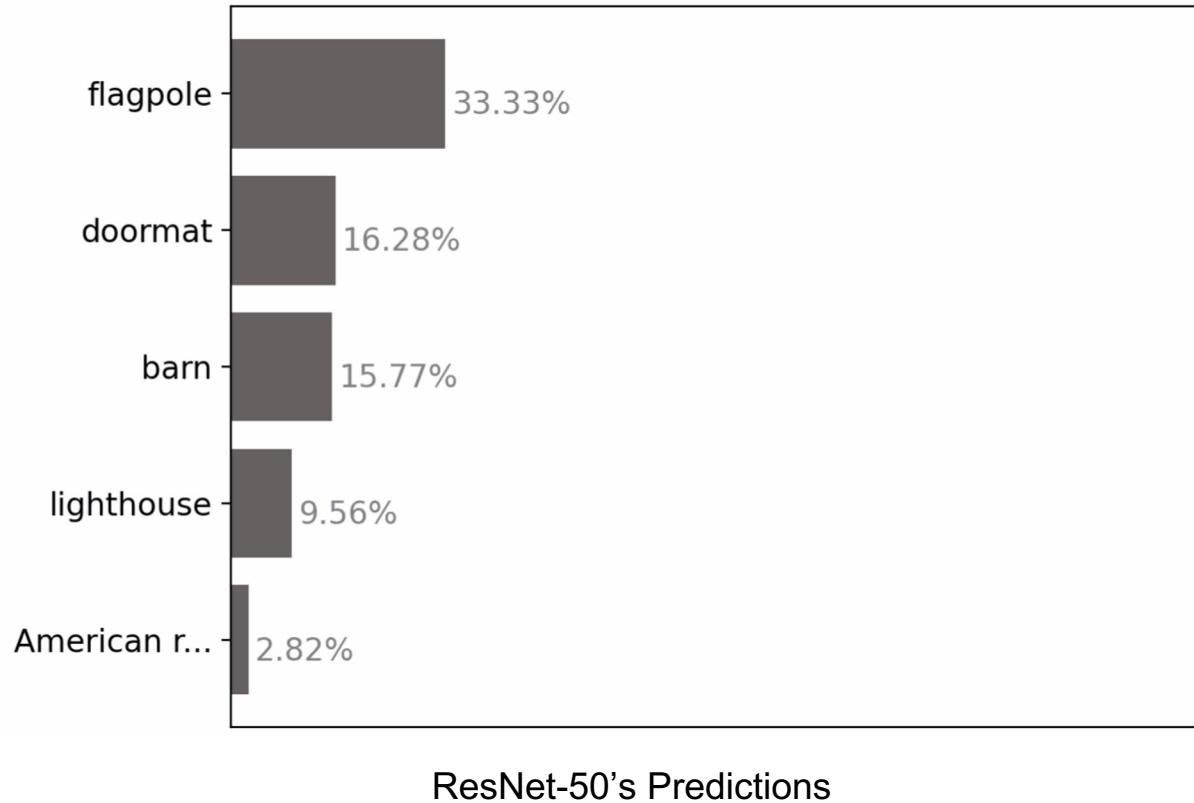


# Continuous Zoom Increases The Probability Of The Correct Class

Zoom Level: 256



Image source: ImageNet-A



***Zooming in allows us to see patterns more clearly and eliminate distractions.***



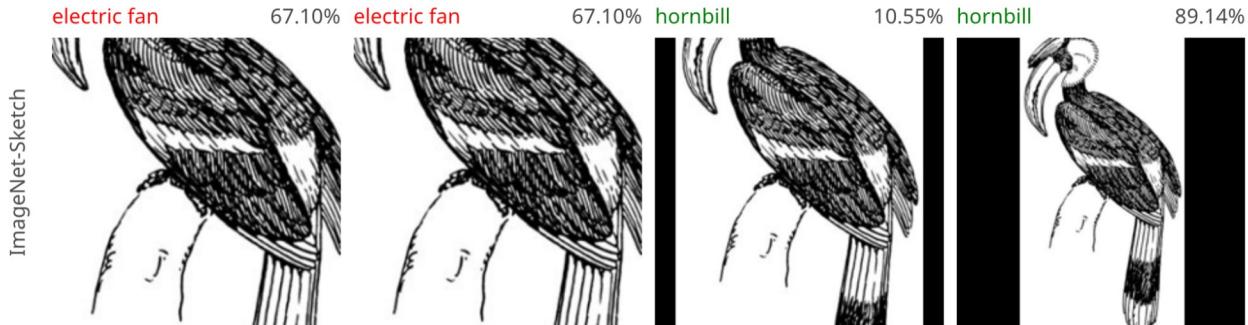
Image source: ImageNet-A and Sketch

Predictions from ResNet-50 and ViT-B/32 classifiers

***Zooming in allows us to see patterns more clearly and eliminate distractions.***



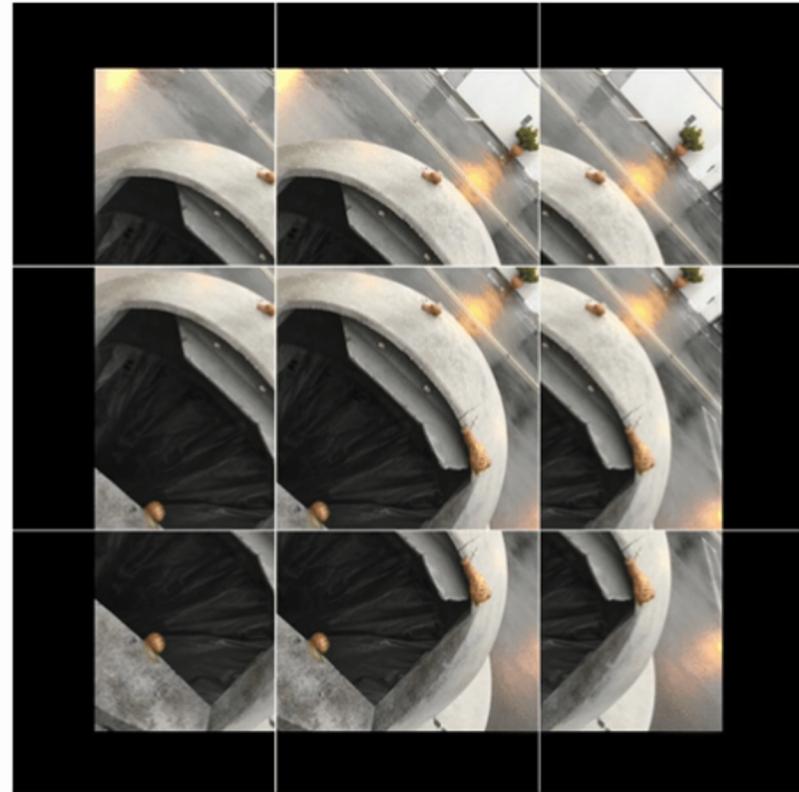
***Zooming out enables us to have a better view of the entire object.***



# Defining Zoom Transforms

# Our Zoom Operation Consists Of *Resize* And *Crop* Operations

- We test **36** different zoom levels ( $T$ )
- $T$  varies from **10** to **1024** pixels
- An image will be transformed into **324** different "**zoomed**" versions



If One Out Of 324 Zoomed Versions Is Classified Correctly, We Deem The Image Classifiable

**Upper-bound Accuracy:** The ratio of a dataset that can be correctly classified after  $N = 36$  or 324 zoom attempts



...



bald eagle

If One Out Of 324 Zoomed Versions Is Classified Correctly, We Deem The Image Classifiable

**Upper-bound Accuracy:** The ratio of a dataset that can be correctly classified after  $N = 36$  or 324 zoom attempts

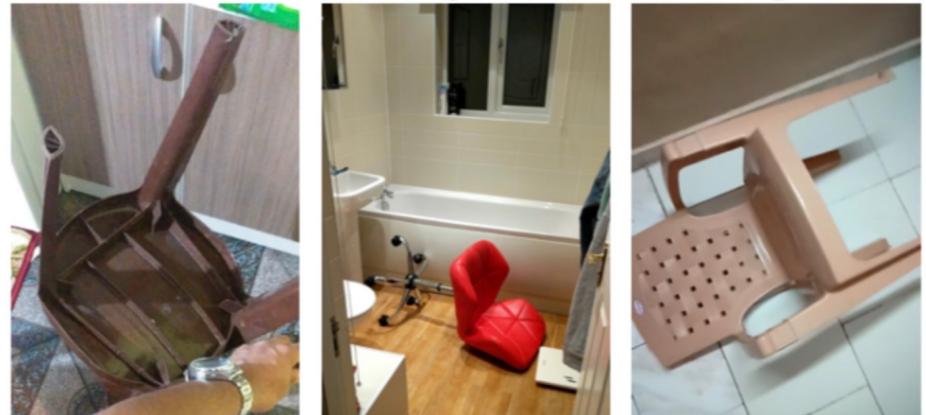


This image is “*classifiable*”

# Significant Gap Between 1-crop and Maximum Possible Accuracy

	IN	ReaL	IN+ReaL	IN-A	IN-R	IN-S	ON
<b>(a) Standard top-1 accuracy based on <math>N = 1</math> crop</b>							
AlexNet	56.16	62.67	61.76	1.75	21.10	10.05	14.23
VGG-16	71.37	78.90	78.52	2.69	26.98	16.78	28.32
ResNet-18	69.45	76.94	76.47	1.37	32.14	19.41	27.59
ResNet-50	75.75	82.63	82.97	0.21	35.39	22.91	36.18
ViT-B/32	75.75	81.89	82.59	9.64	41.29	26.83	30.89
<b>CLIP-ViT-L/14</b>	75.03	80.68	81.95	71.28	87.74	58.23	66.32
<b>(b) Upper-bound accuracy using <math>N = 36</math> crops</b>							
Random	3.60	3.60	3.60	18.00	18.00	3.60	31.85
AlexNet	85.19	90.30	89.74	31.37	47.04	24.40	49.17
VGG-16	92.30	96.08	95.81	46.69	52.86	34.34	62.94
ResNet-18	92.08	95.97	95.73	47.48	58.85	37.91	63.08
ResNet-50	94.46	97.36	97.40	55.68	61.42	41.71	69.60
ViT-B/32	95.05	97.61	97.88	68.43	68.77	49.10	70.30
<b>CLIP-ViT-L/14</b>	94.19	97.32	97.56	97.16	98.60	83.77	89.59

# Positional Bias: Maximum Possible Accuracy Is Higher At Center



*ObjectNet images taken by mobile phones*

94.65 (-2.12)	95.92 (-0.85)	94.94 (-1.83)
95.58 (-1.19)	96.77	95.91 (-0.86)
94.53 (-2.24)	95.82 (-0.95)	94.82 (-1.95)

ImageNet-Real

22.52 (-23.97)	27.61 (-18.88)	22.31 (-24.18)
27.57 (-18.92)	46.49	26.57 (-19.92)
21.17 (-25.32)	26.77 (-19.72)	21.59 (-24.90)

ImageNet-A

57.09 (-5.43)	59.60 (-2.92)	57.19 (-5.33)
59.49 (-3.03)	62.52	59.62 (-2.90)
57.55 (-4.97)	60.28 (-2.24)	57.59 (-4.93)

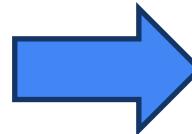
ImageNet-R

48.42 (-20.88)	49.84 (-19.46)	48.55 (-20.75)
62.25 (-7.05)	69.30	62.53 (-6.77)
46.81 (-22.49)	47.96 (-21.34)	46.85 (-22.45)

ObjectNet

# Exploiting **ImageNet-A** and **ObjectNet** Center Bias

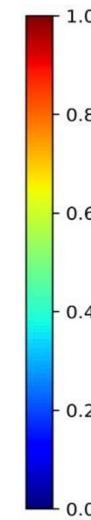
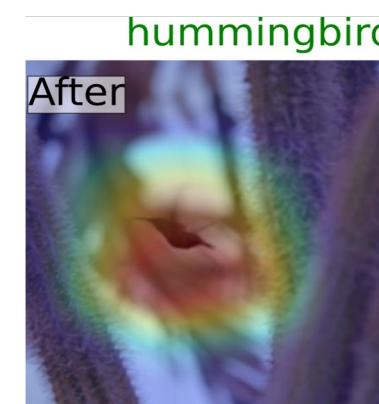
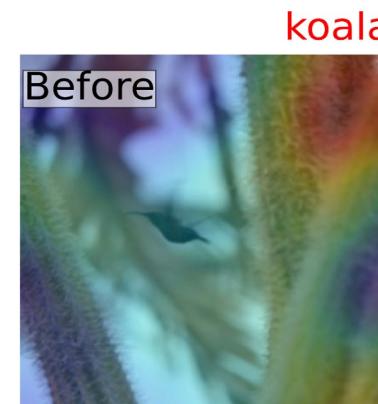
Simply **center-cropping** the image improve the accuracy of classifiers on ImageNet and ObjectNet.



**4 to 14** % points

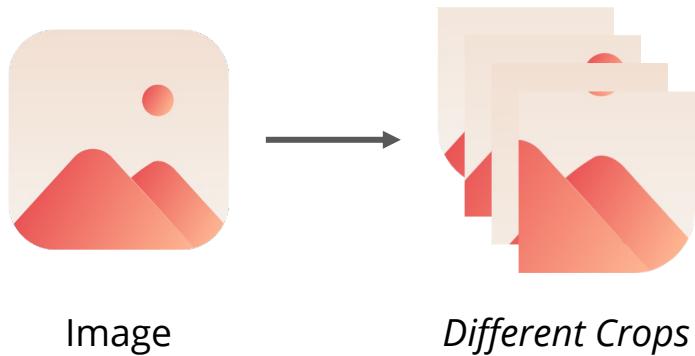
**ResNet-50** accuracy on **ImageNet-A**: ~0%

# ResNet-50 (GradCAM) focuses on object after MEMO update

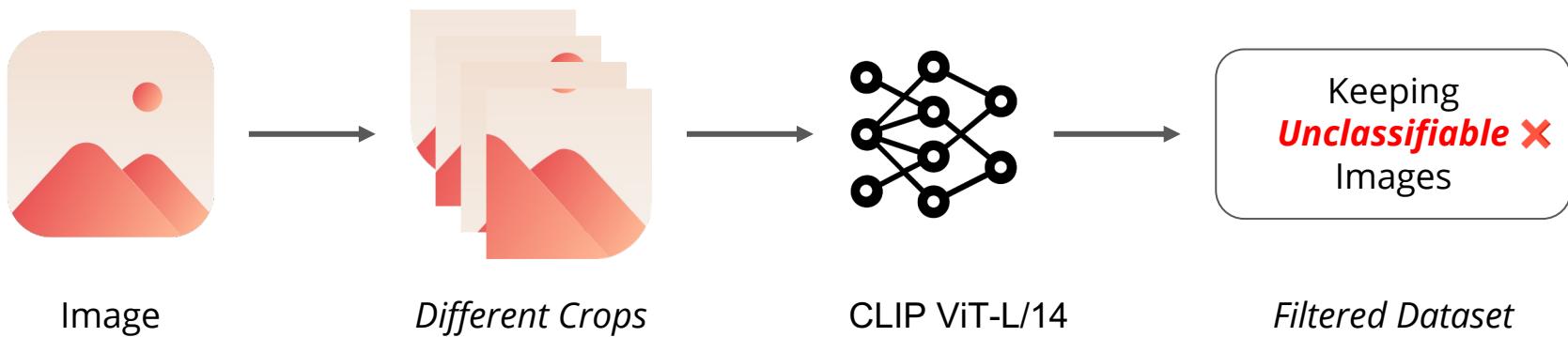


# Introducing *ImageNet-Hard*

# Creating ImageNet-Hard



# Creating ImageNet-Hard



Image

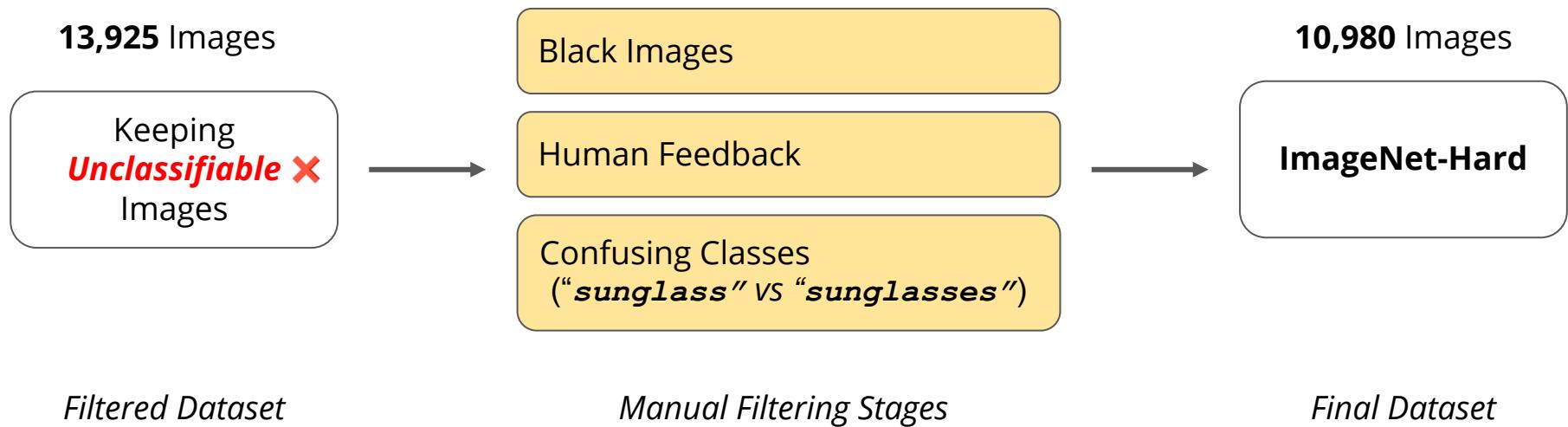
*Different Crops*

CLIP ViT-L/14

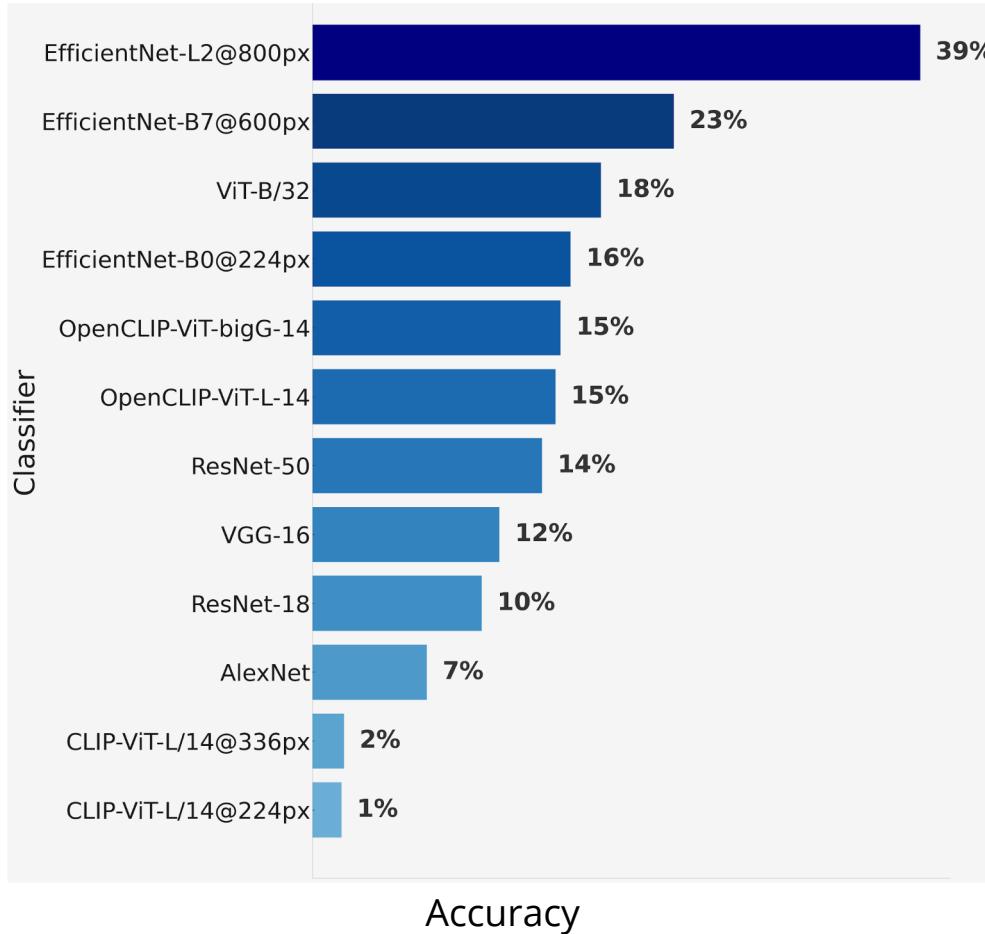
*Filtered Dataset*

**Unclassifiable:** None of the 324 crops are correctly classified

# Imagenet-Hard Consists Of 10,980 Unclassifiable Images Even After 324 Zoom Attempts



# The Best model only scores **39%** on **Imagenet-Hard**

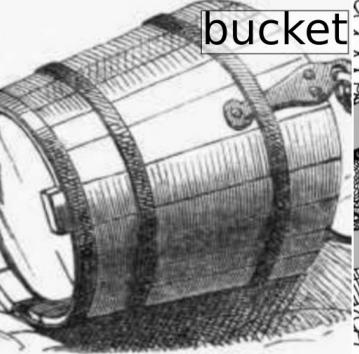


beaker



cup

barrel



bucket

rock beauty



clownfish

cassette



tape player

ruler



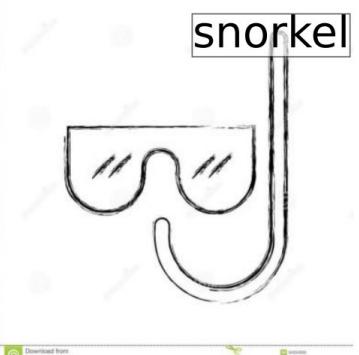
slide rule

weighing scale



tray

hook



snorkel

quill

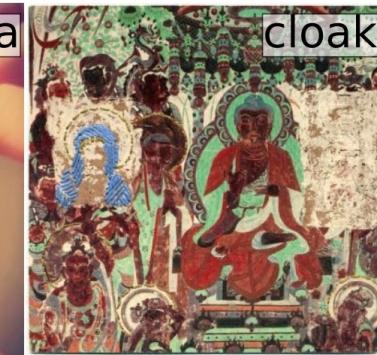


tray

plectrum



jigsaw puzzle



cloak

predicted label by EfficientNet-L2@800px

groundtruth label

# GlitchBench

Testing LMMs on game glitch detection

<https://glitchbench.github.io/>

The video game industry is the largest and fastest growing sector of the entertainment industry, surpassing the film and music industries.



- Music: 26.2 B
- Movie: 26B

**184 billion  
USD** in 2022

# What is wrong with this scene?



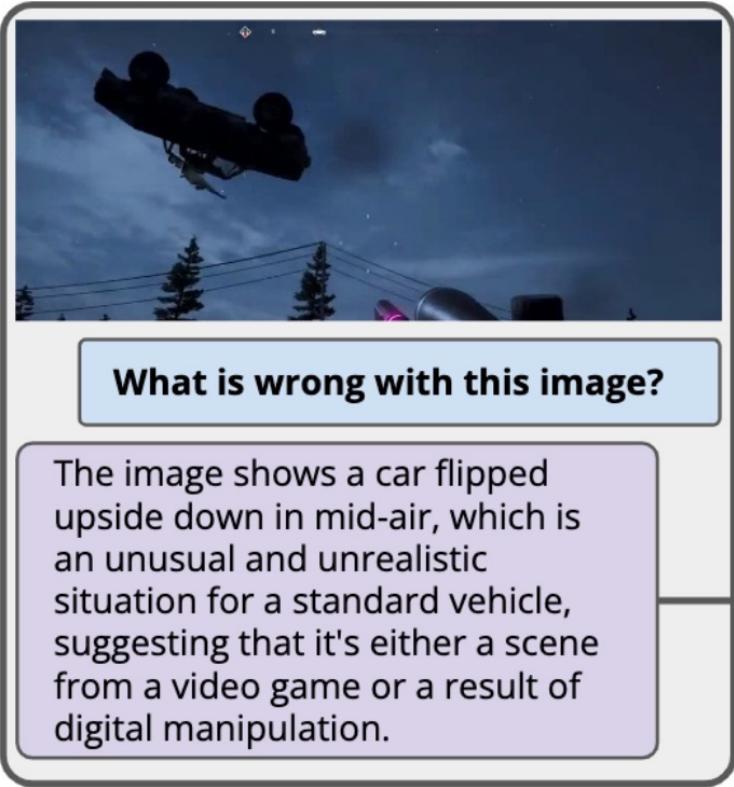
# What is wrong with this scene?



**Ground truth:** It is raining inside the building

**Question:** What is wrong with this image?

**GPT-4V:** There's nothing inherently "wrong" with this image; however, if you're referring to inconsistencies or peculiarities that might not align with the established Batman canon, the red Batman suit is unusual since Batman is traditionally known in black or very dark colors. The use of red could be an artistic choice for the game, an alternate suit available within the game's universe, or indicative of a different character who takes on the Batman mantle.



**Ground Truth:**

A car is upside down in the air.

 **Judge**  
(Llama-2)

**Yes**

**No**

Figure 3. To evaluate a model's response, we ask a judge (the Llama-2-70b-Chat model) to compare it semantically with the ground truth.



*Two people driving an invisible car*



*A rifle floating in the air*



<https://www.reddit.com/r/GamePhysics>



*The gun in the hand is missing*



*The table cover has a placeholder texture*



*Two people driving an invi:*



*The gun in the hand is m*

Model	▲	What is unusual about this image?
GPT-4V(ision)		56.8
LLaVA-1.5-13B		36.3
LLaVA-1.5-7B		35.2
Long-SPHINX		25.3
InstructBLIP-7B		25.3
OtterHD		24.8
InstructBLIP-13B		21.9
Qwen-VL-Chat		21.2
SPHINX		19.2
MiniGPT4v2		19.1
Fuyu-8B		8.4



what is wrong in this image?



### ChatGPT

The image shows what appears to be a tense scene with a large, armored vehicle labeled "GCPD," which stands for Gotham City Police Department, a fictional law enforcement agency from the Batman universe. Two individuals are seated on the ground with their backs against the vehicle. It's hard to say definitively what is "wrong" in the image without context, as it might be a scene from a game, movie, or another form of media related to the Batman franchise. If you're looking for inconsistencies or errors, nothing immediately stands out as incorrect; it all depends on the context of the scene depicted.

# Long-term Research Goals

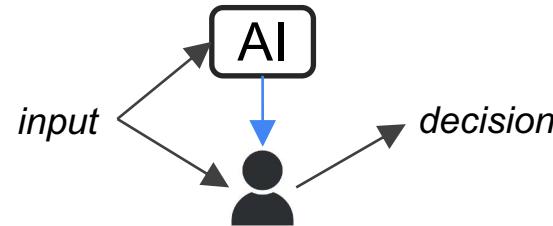
## Trustworthy AI

1. Build AIs that are accurate in edge cases  
(and common cases)

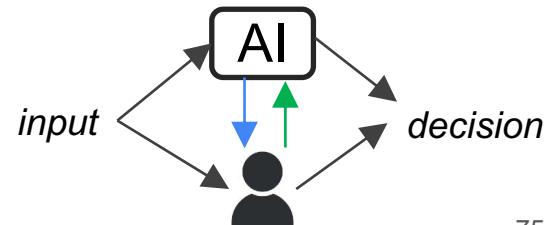


## Explainable AI

2. Build AIs that maximize human-AI team accuracy



3. Build AIs that humans can debug and edit  
(AI's decision-making process)



# Long-term Research Goals

## Trustworthy AI

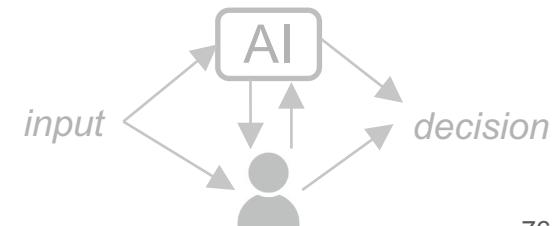
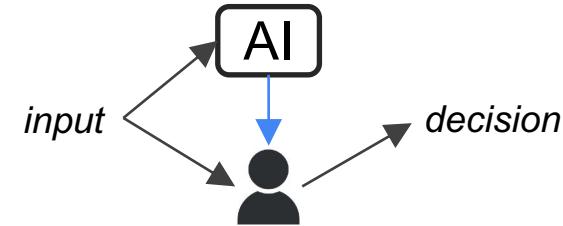
1. Build AIs that are accurate in edge cases  
(and common cases)

## Explainable AI

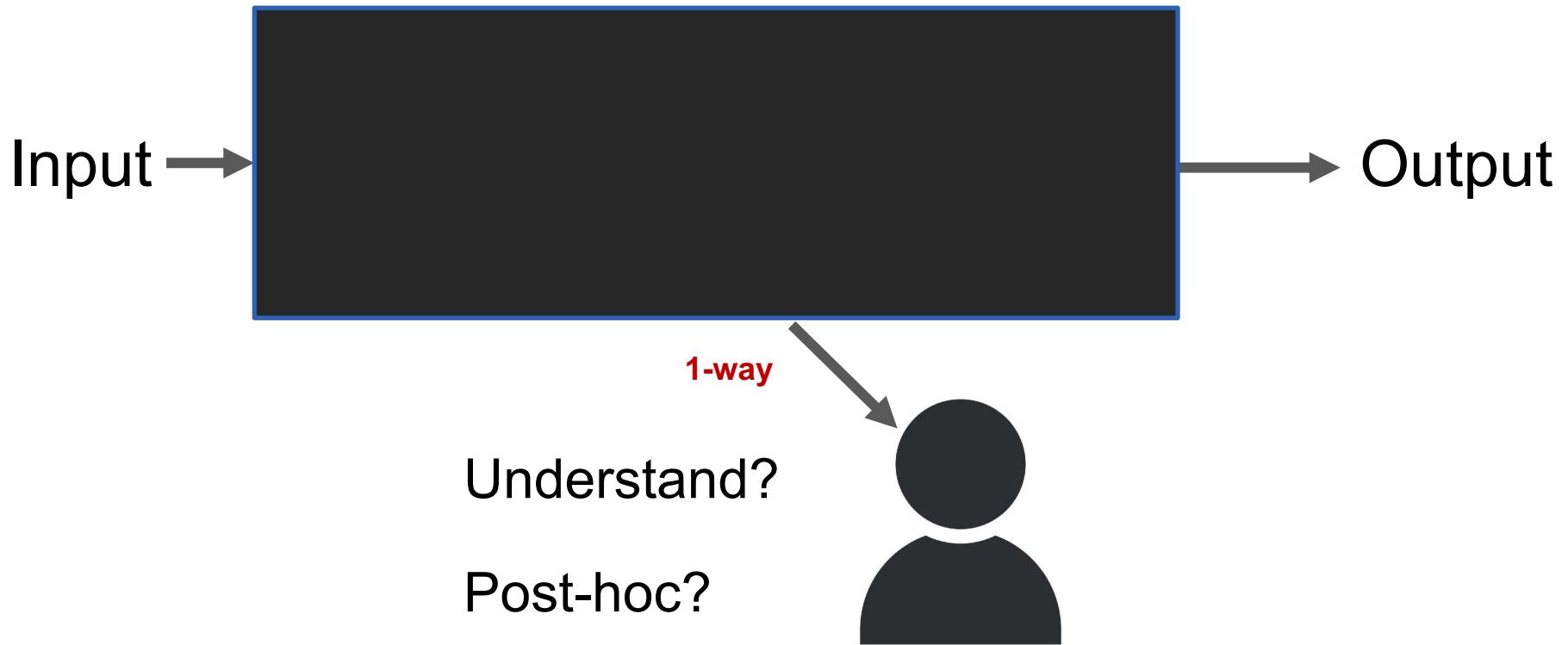
2. Build AIs that maximize human-AI team accuracy

3. Build AIs that humans can debug and edit  
(their decision-making process)

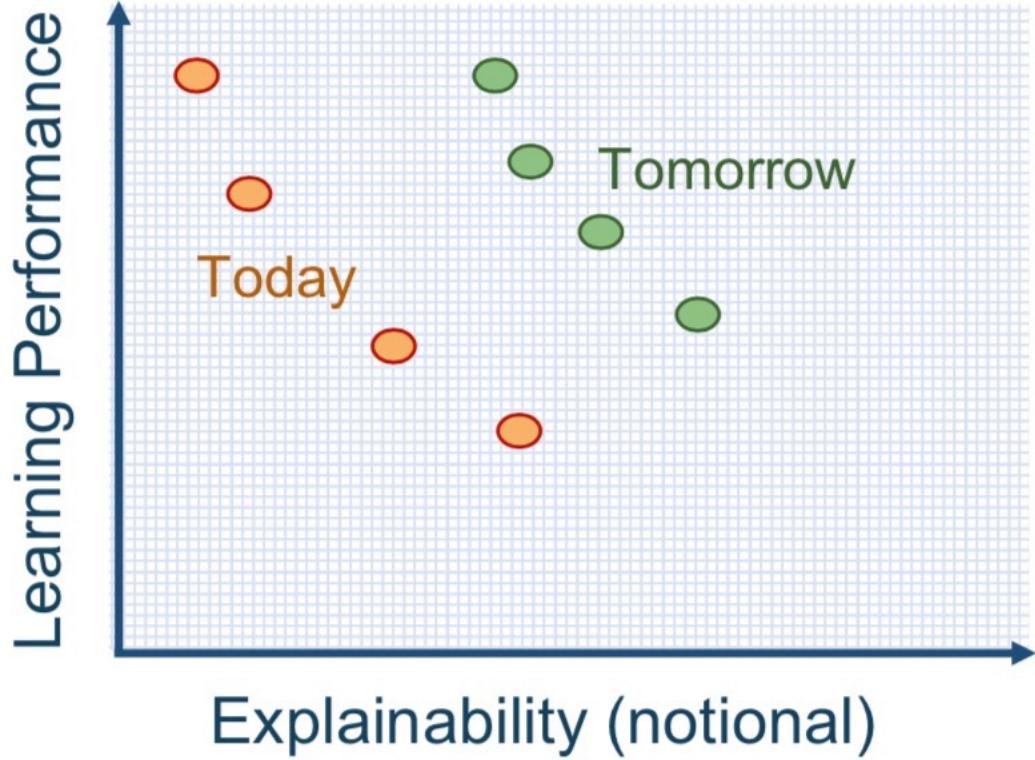
# Part 2



# Existing AI and XAI



# Performance vs. Explainability



DARPA XAI 2016-2021

<https://www.darpa.mil/attachments/XAIProgramUpdate.pdf>

## Measure of Explanation Effectiveness

### User Satisfaction

- Clarity of the explanation (user rating)
- Utility of the explanation (user rating)

### Mental Model

- Understanding individual decisions
- Understanding the overall model
- Strength/weakness assessment
- 'What will it do' prediction
- 'How do I intervene' prediction

### Task Performance

- Does the explanation improve the user's decision, task performance?
- Artificial decision tasks introduced to diagnose the user's understanding

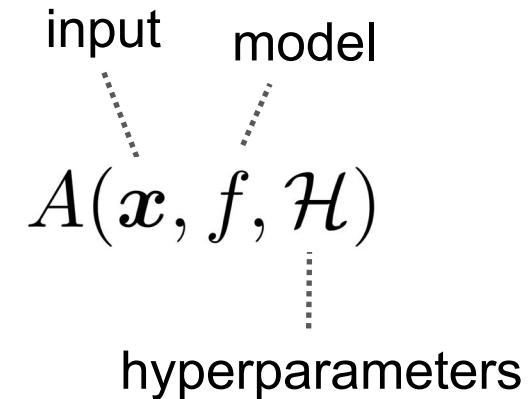
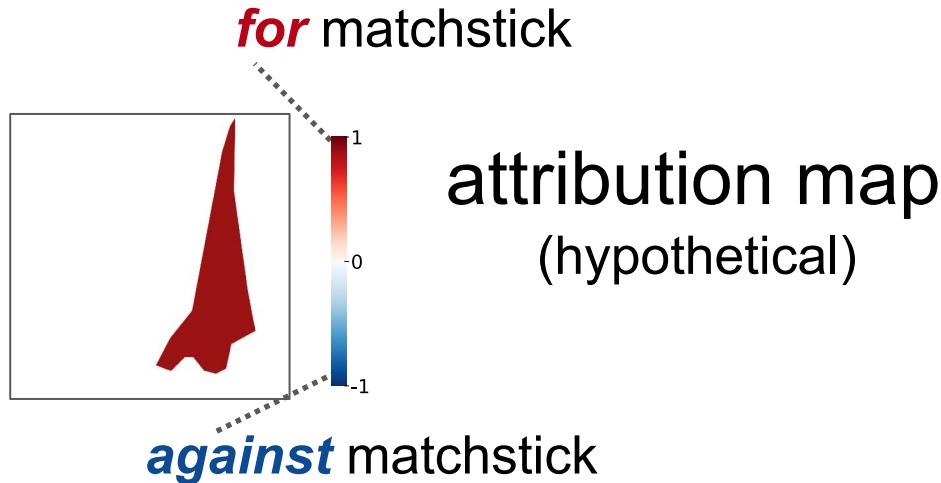
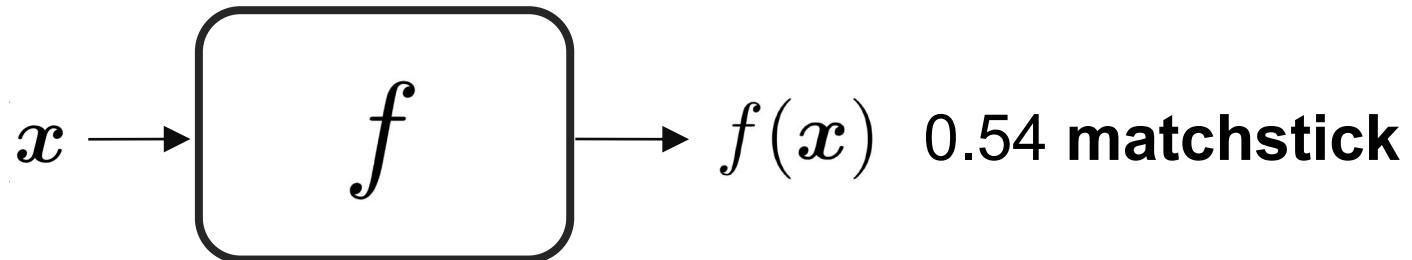
### Trust Assessment

- Appropriate future use and trust

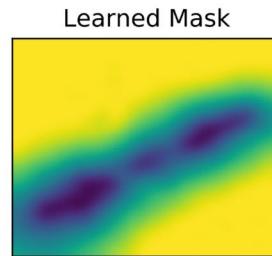
### Correctability (Extra Credit)

- Identifying errors
- Correcting errors
- Continuous training

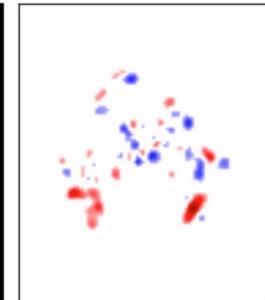
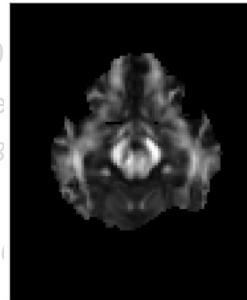
# Attribution maps: What input features cause 0.54 matchstick?



- Deconvnet: Visualizing and understanding convolutional networks. Zeiler et al. 2014
- Guided-backprop: Striving for simplicity: The all convolutional net. Springenberg et al. 2015
- Integrated Gradients: Axiomatic Attribution for Deep Networks. Sundararajan et al. 2018
- CAM: Learning Deep Features for Discriminative Localization. Zhou et al. 2016
- LIME: Why should i trust you?: Explaining the predictions of any classifier. Ribeiro et al. 2016
- SmoothGrad: removing noise by adding noise. Smilkov et al. 2017
- MP: Interpretable Explanations of Black Boxes by Meaningful Perturbation. Fong et al. 2017
- SHAP: A Unified Approach to Interpreting Model Predictions. Lundberg et al. 2017
- PDA: Visualizing deep neural network decisions: Prediction difference analysis. Zintgraf et al. 2017
- Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Selvaraju et al. 2017
- Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. Chattopadhyay et al. 2017
- LRP: Beyond saliency: understanding convolutional neural networks from saliency prediction on layer-wise relevance propagation
- DeepLIFT: Learning important features through propagating activation differences. Shrikumar et al. 2017
- RISE: Randomized Input Sampling for Explanation of Black-box Models. Petsiuk et al. 2018
- FIDO: Explaining image classifiers by counterfactual generation. Chang et al. 2019
- Expected Gradients: Learning Explainable Models Using Attribution Priors. Erion et al. 2019
- FG-Vis: Interpretable and Fine-Grained Visual Explanations for Convolutional Neural Networks. Wagner et al. CVPR 2019
- Understanding Deep Networks via Extremal Perturbations and Smooth Masks. Fong et al. ICCV 2019
- MP-G: Removing input features via a generative model to explain their attributions to classifier's decisions. Agarwal et al. 2020



Zeiler et al. 2014  
Springenberg et al. 2015  
Jain et al. 2018  
2016  
. Ribeiro et al. 2016



## Smooth Natural images Fong et al. 2017

MP: Interpreting Linear Models via Indirect Perturbation. Fong et al. 2017

SHAP: A Unified Approach to Interpreting Model Predictions. Lundberg et al. 2017

PDA: Visualizing deep neural network decisions: Prediction difference analysis. Fong et al. 2017

From: johnchad@triton.unm.edu (jchadwick)

Subject: Another request for Darwin Fish

Organization: University of New Mexico, Albuquerque

Lines: 11

NNTP-Posting-Host: triton.unm.edu

FIDO: Explainable Generation. Ribeiro et al. 2016

Expected Gradients: Learning Explainable Models Using Attribution Priors.

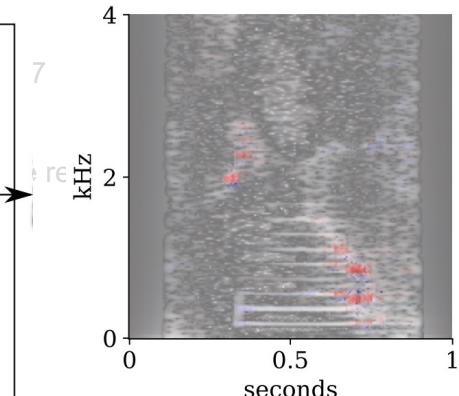
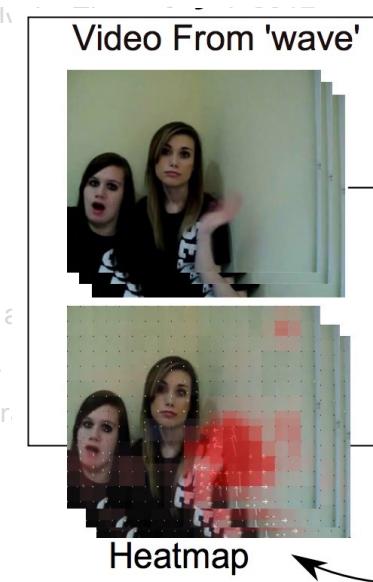
FG-Vis: Interpretable and Fine-Grained Visual Explanations for Convolutional Neural Networks.

Understanding Deep Networks via Extremal Perturbations. Srinivasan et al. 2017

MP-G: Removing input features via a generative model. Agarwal et al. 2019

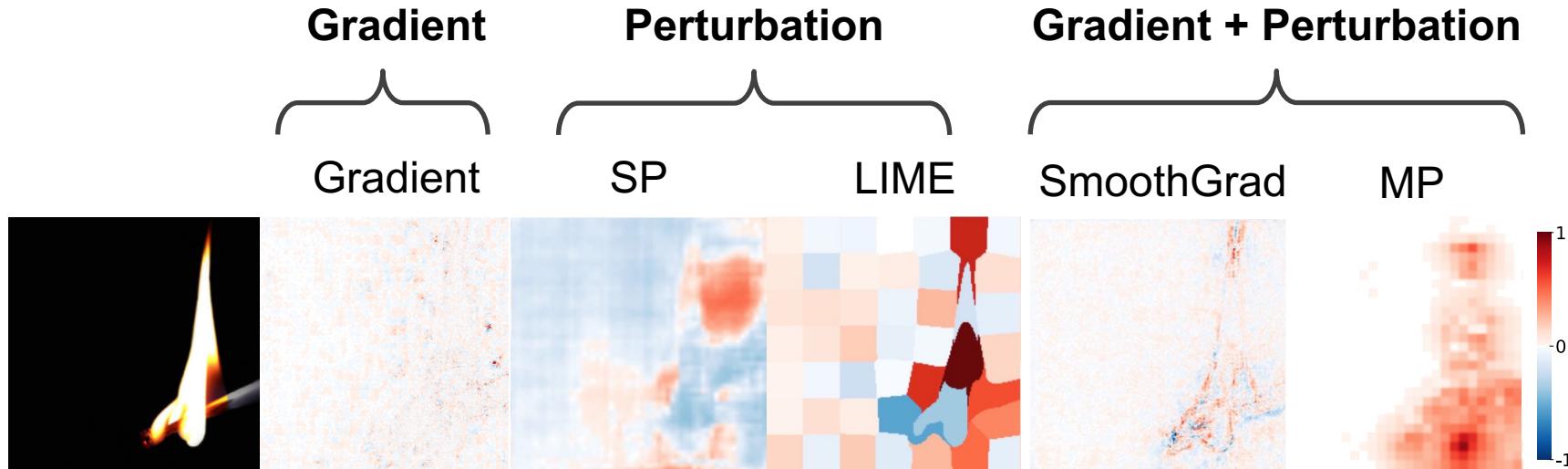
...  
Srinivasan et al. 2017

## Videos



## Audio

Agarwal et al. 2019



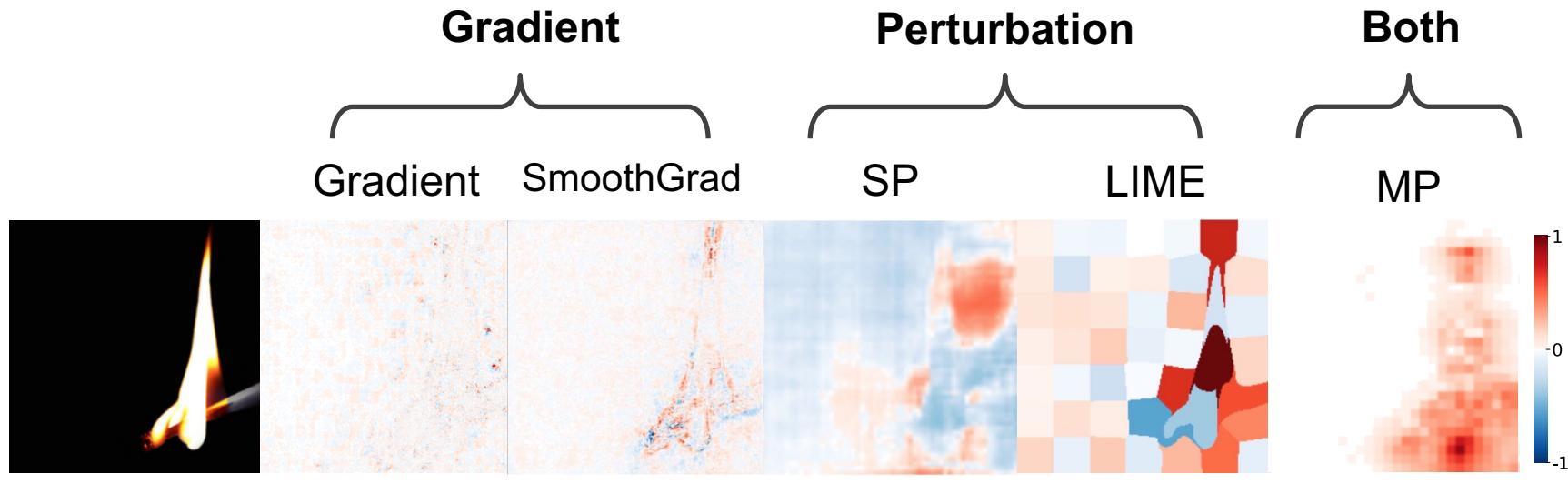
0.54 matchstick

Zeiler & Fergus 2014 Ribeiro et al. 2016 Smilkov et al. 2017 Fong & Vedaldi 2017

Are these explanations correct and reliable?

# Problems with most attribution methods (“saliency methods”)

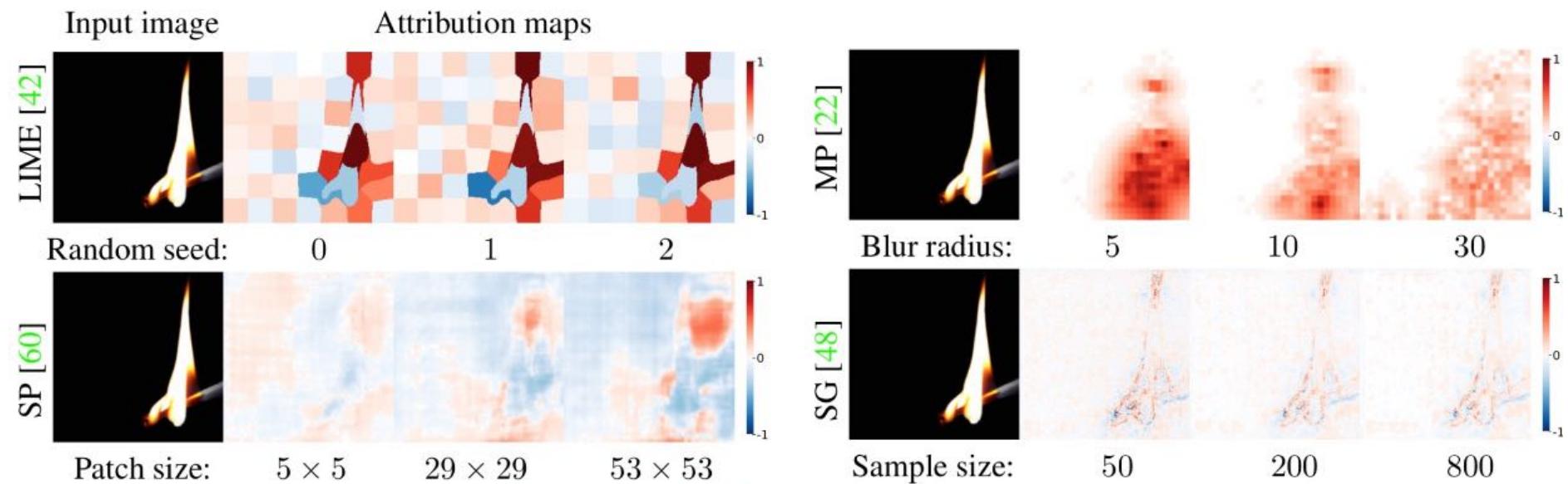
1. After-the-fact (posthoc) explanations attempting to explain a black-box AI.



0.54 matchstick

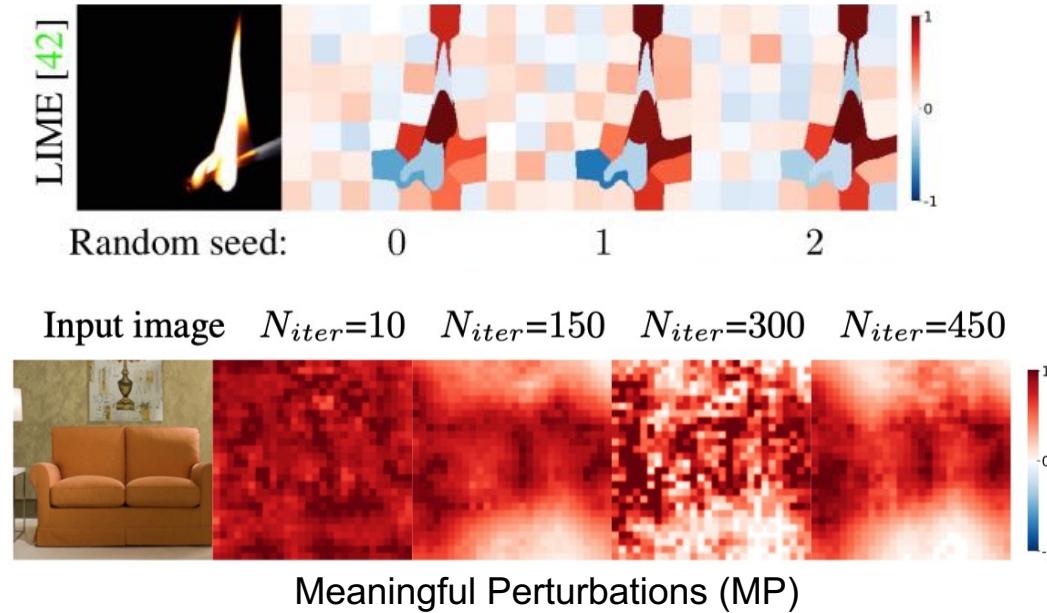
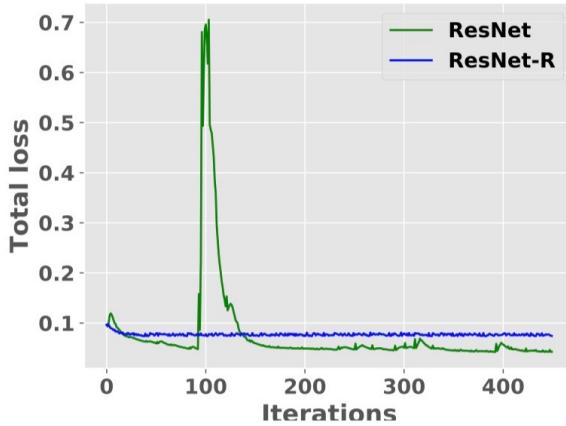
# Problems with most attribution methods (“saliency methods”)

1. After-the-fact explanations attempting to explain a black-box AI.
2. Sensitive to (many) hyperparameters (some of which are arbitrarily tuned)



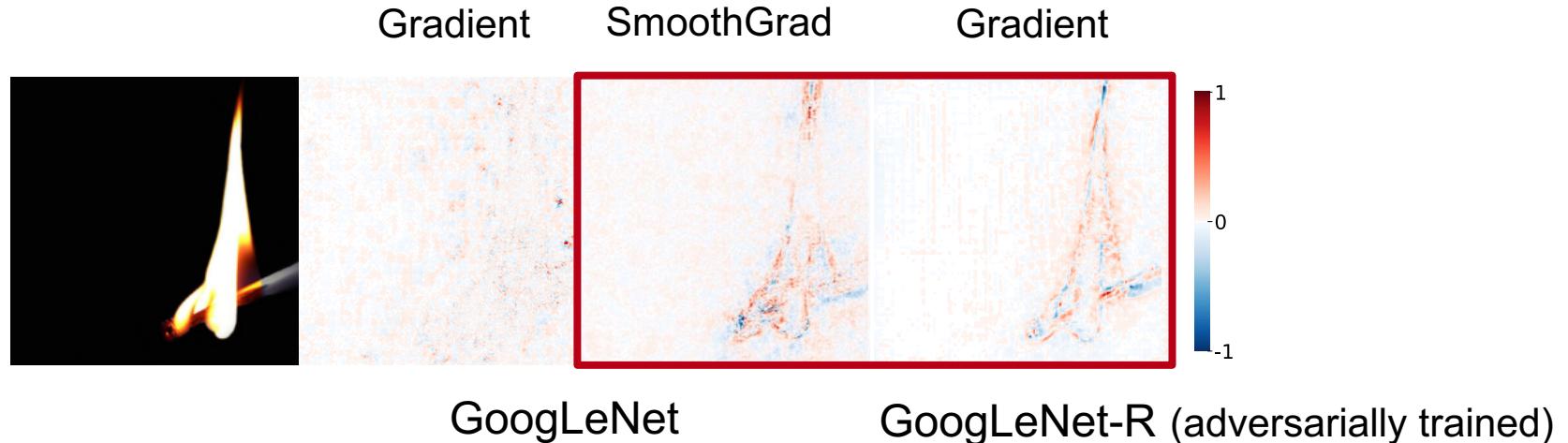
# Problems with most attribution methods (“saliency methods”)

1. After-the-fact explanations attempting to explain a black-box AI.
2. Sensitive to (many) hyperparameters (some of which are arbitrarily tuned)
3. Hard to reproduce (due to iterative optimization & random sampling)



# Problems with most attribution methods (“saliency methods”)

1. After-the-fact explanations attempting to explain a black-box AI.
2. Sensitive to (many) hyperparameters (some of which are arbitrarily tuned)
3. Hard to reproduce (due to iterative optimization & random sampling)
4. Can be misleading



# Problems with most attribution methods (“saliency methods”)

1. After-the-fact explanations attempting to explain a black-box AI.
2. Sensitive to (many) hyperparameters (some of which are arbitrarily tuned)
3. Hard to reproduce (due to iterative optimization & random sampling)
4. Can be misleading
5. Highlights the same main object when AI is either correct or wrong

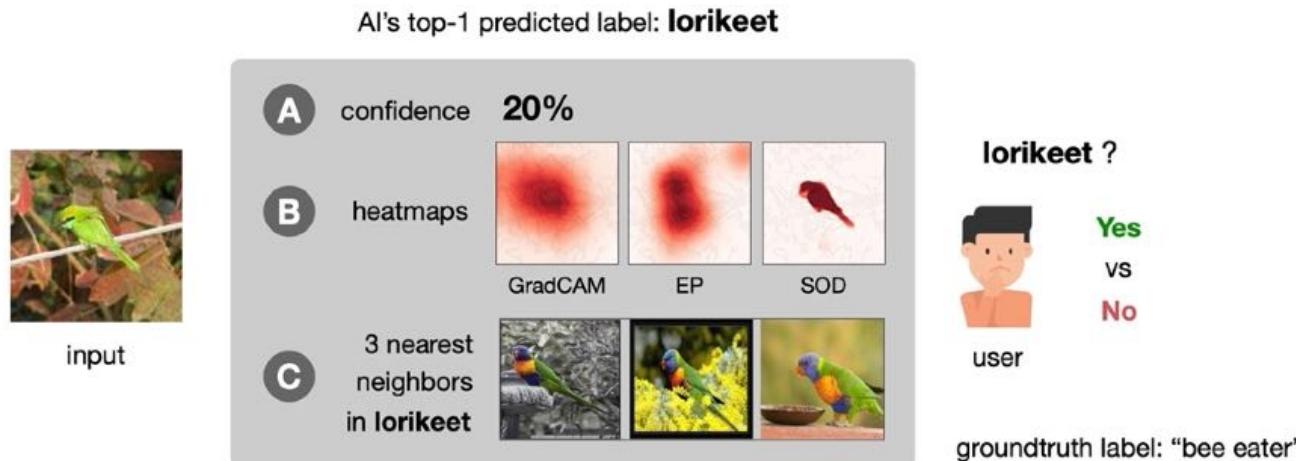


Top Label: jeep, landrover  
Top Label Probability: 0.8962

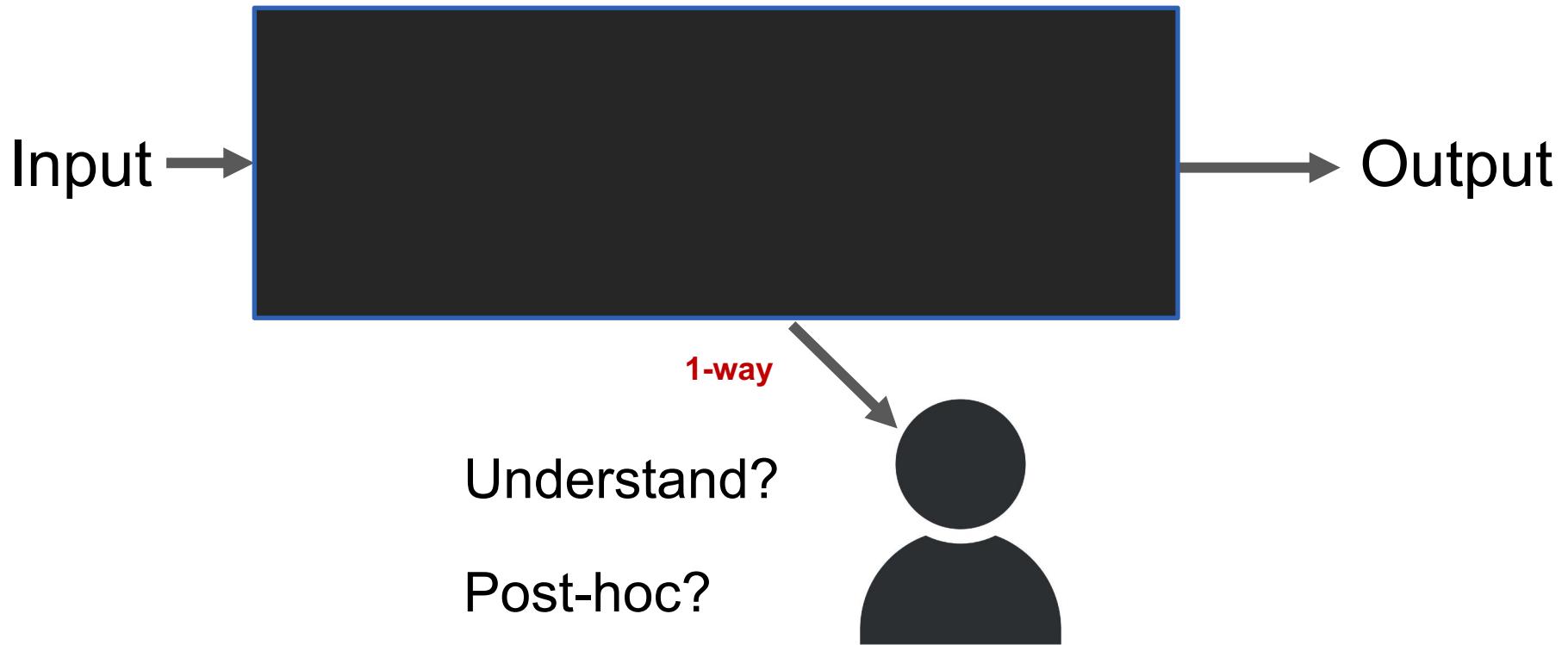
Top Label: snowplow, snowplough  
Top Label Probability: 0.9639

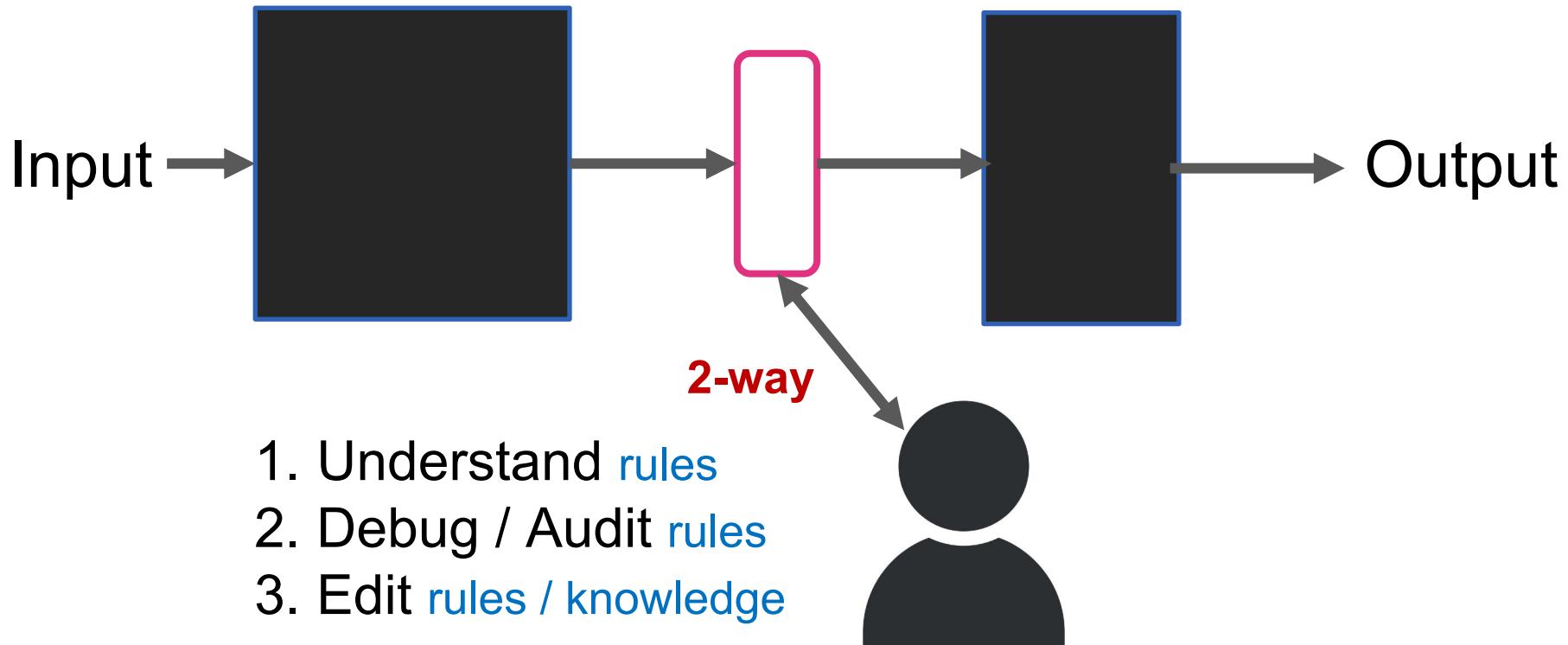
# Problems with most attribution methods (“saliency methods”)

1. After-the-fact explanations attempting to explain a black-box AI.
2. Sensitive to (many) hyperparameters (some of which are arbitrarily tuned)
3. Hard to reproduce (due to iterative optimization & random sampling)
4. Can be misleading
5. Highlights the same main object when AI is either correct or wrong



# Existing AI and XAI



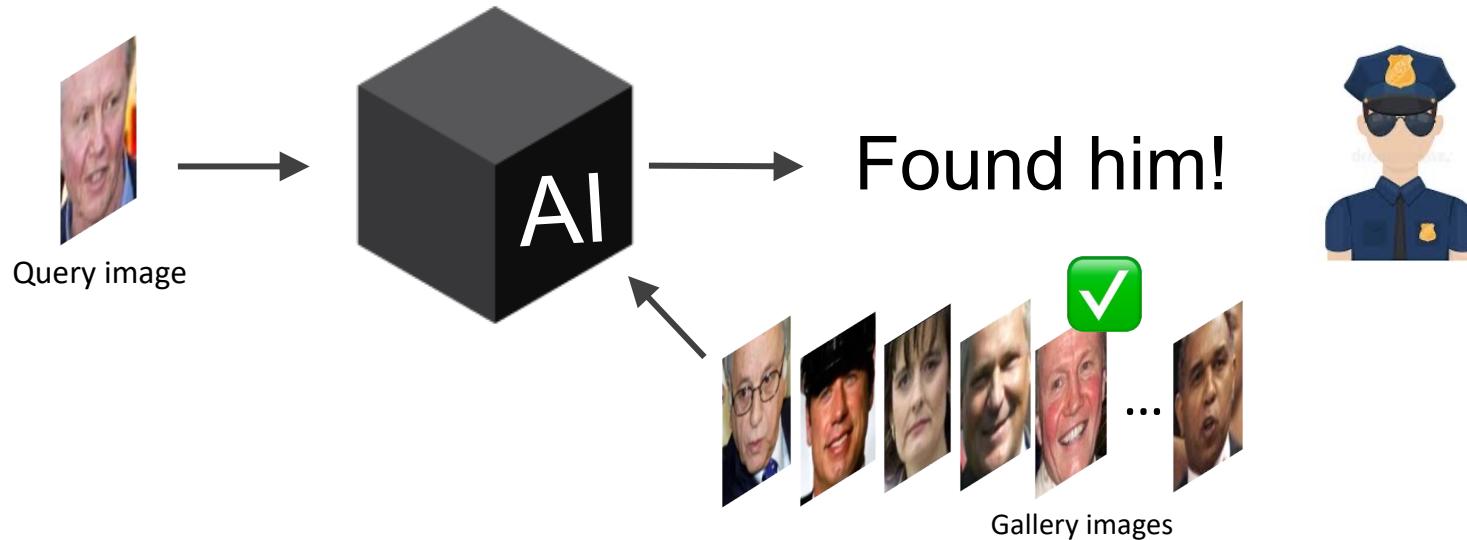


# Solution: “Slow-thinking” System 2, Explainable AIs

- First explain, Then decide
- Harness external knowledge-bases to make informed decisions

An example in **Face Identification**

**Challenge:** Zero-shot, fine-grained classification



# The new lawsuit that shows facial recognition is officially a civil rights issue

Robert Williams, who was wrongfully arrested because of a faulty facial recognition match, is asking for the technology to be banned.

by **Tate Ryan-Mosley**

April 14, 2021

**On January 9, 2020, Detroit police drove to the suburb of Farmington Hill and arrested Robert Williams in his driveway while his wife and young daughters looked on. Williams, a Black man, was accused of stealing watches from Shinola, a luxury store. He was held overnight in jail.**

## *Wrongfully Accused by an Algorithm*

In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

# Amazon's Face Recognition Falsely Matches Members of Congress With Mugshots



By [Jacob Snow](#), Technology & Civil Liberties Attorney, ACLU of Northern California

JULY 26, 2018 | 8:00 AM

TAGS: [Face Recognition Technology](#), [Surveillance Technologies](#), [Privacy & Technology](#)



Amazon's face surveillance technology is the target of growing opposition nationwide, and today, there are 28 more causes for concern. In a test the ACLU recently conducted of the facial recognition tool, called "Rekognition," the software incorrectly matched 28





Masked

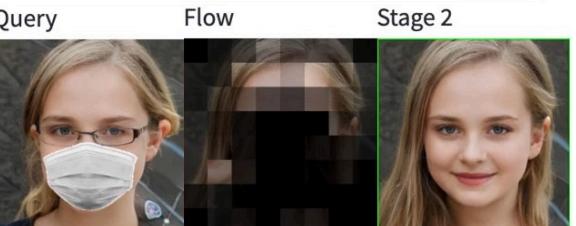
**98.41%**

P@1 on LFW

**39.79%**

P@1 on MLFW

Interactive demo: <https://anhnguyen.me/project/deepface-emd/>

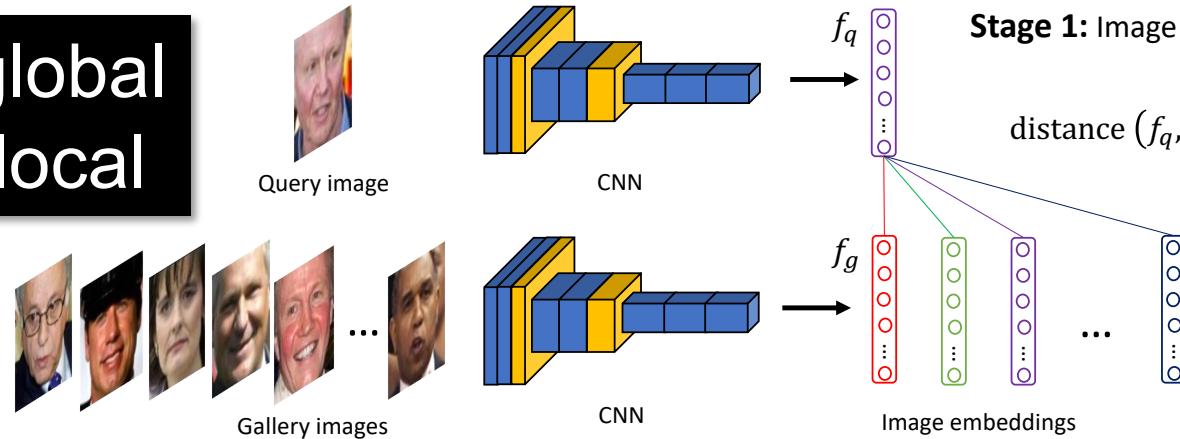


**48.23%**

P@1 on MLFW

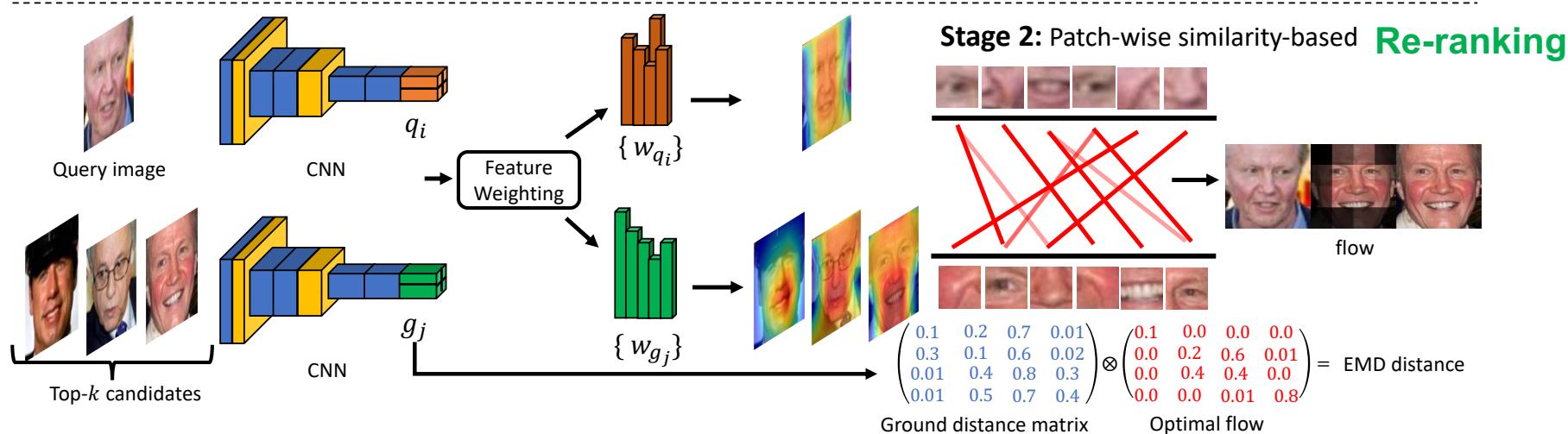
**Ours:** No training  
on masked faces

First, global  
Then, local



Stage 1: Image similarity-based **Ranking**

$$\text{distance } (f_q, f_g) = 1 - \frac{\langle f_q, f_g \rangle}{\|f_q\| \|f_g\|}$$



# Find patch-wise correspondence between two images using optimal transport

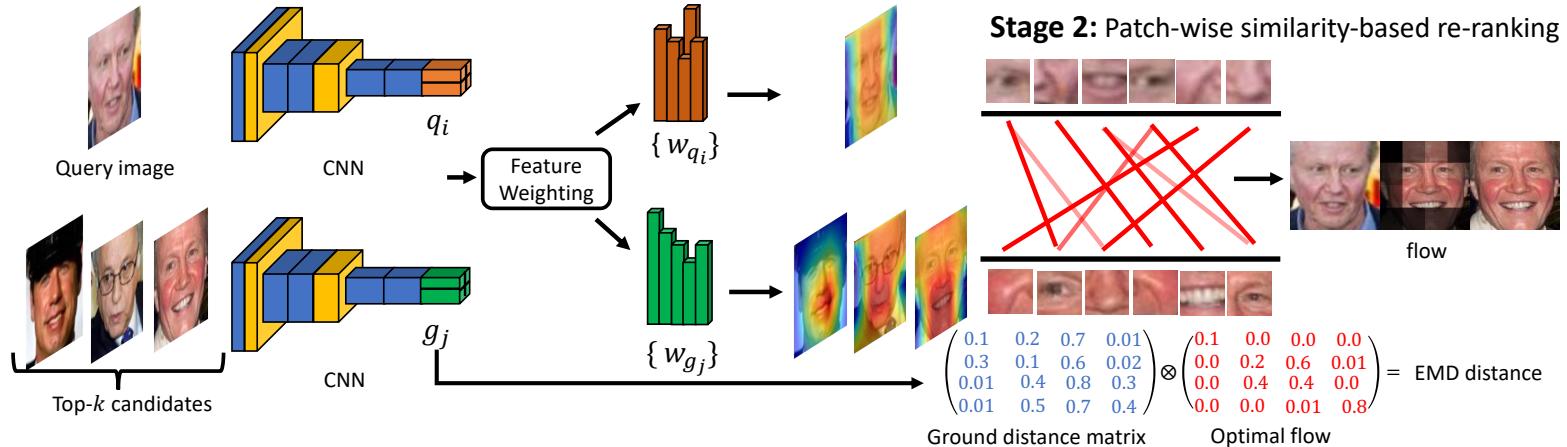
$$\text{s.t. } f_{ij} \geq 0 \quad (2)$$

We want to find an optimal flow  $\mathbf{F}$  that minimizes the following cost function, *i.e.* the sum of weighted pair-wise distances across the two sets of facial features:

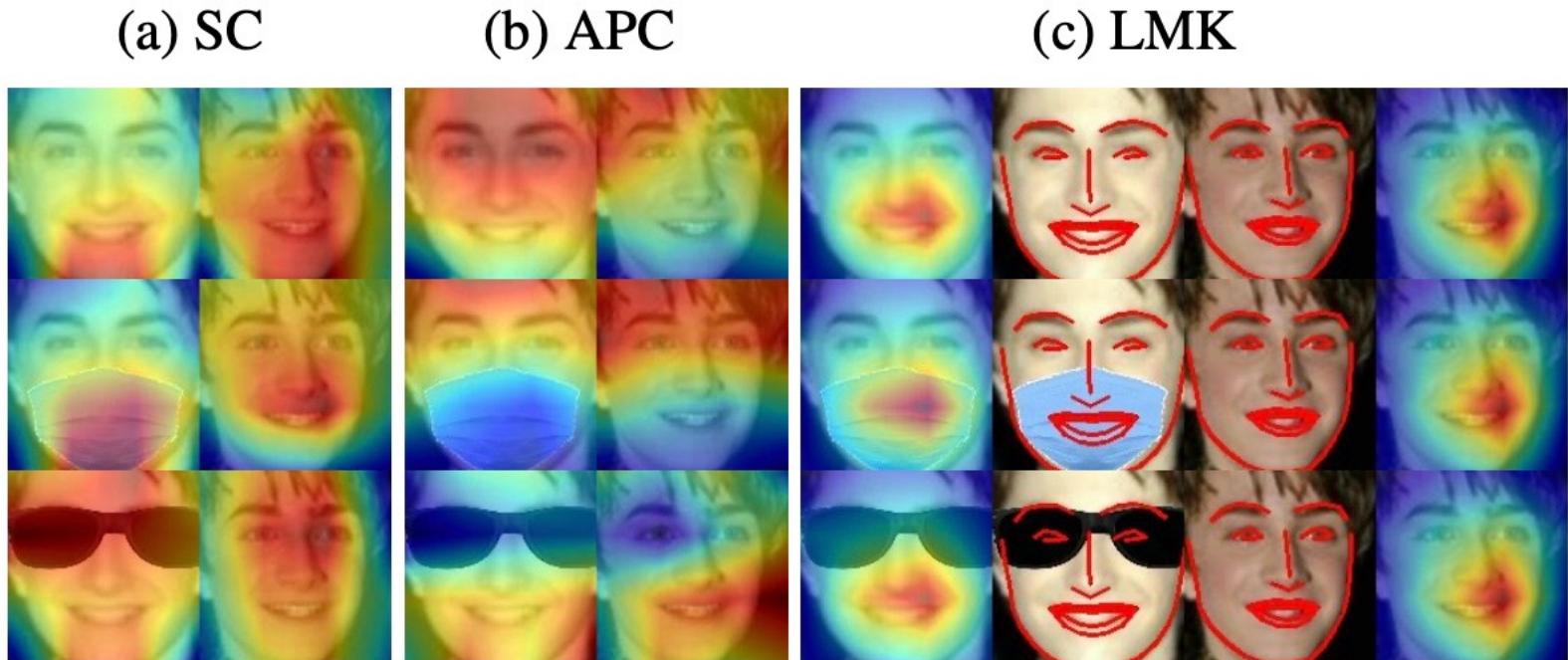
$$\text{COST}(\mathcal{Q}, \mathcal{G}, \mathbf{F}) = \sum_{i=1}^N \sum_{j=1}^N d_{ij} f_{ij} \quad (1)$$

$$\sum_{j=1}^N f_{ij} \leq w_{q_i}, \text{ and } \sum_{i=1}^N f_{ij} \leq w_{g_j}, i, j \in [1, N] \quad (3)$$

$$\sum_{j=1}^N \sum_{i=1}^N f_{ij} = \min \left( \sum_{j=1}^N w_{g_j}, \sum_{i=1}^N w_{q_i} \right). \quad (4)$$

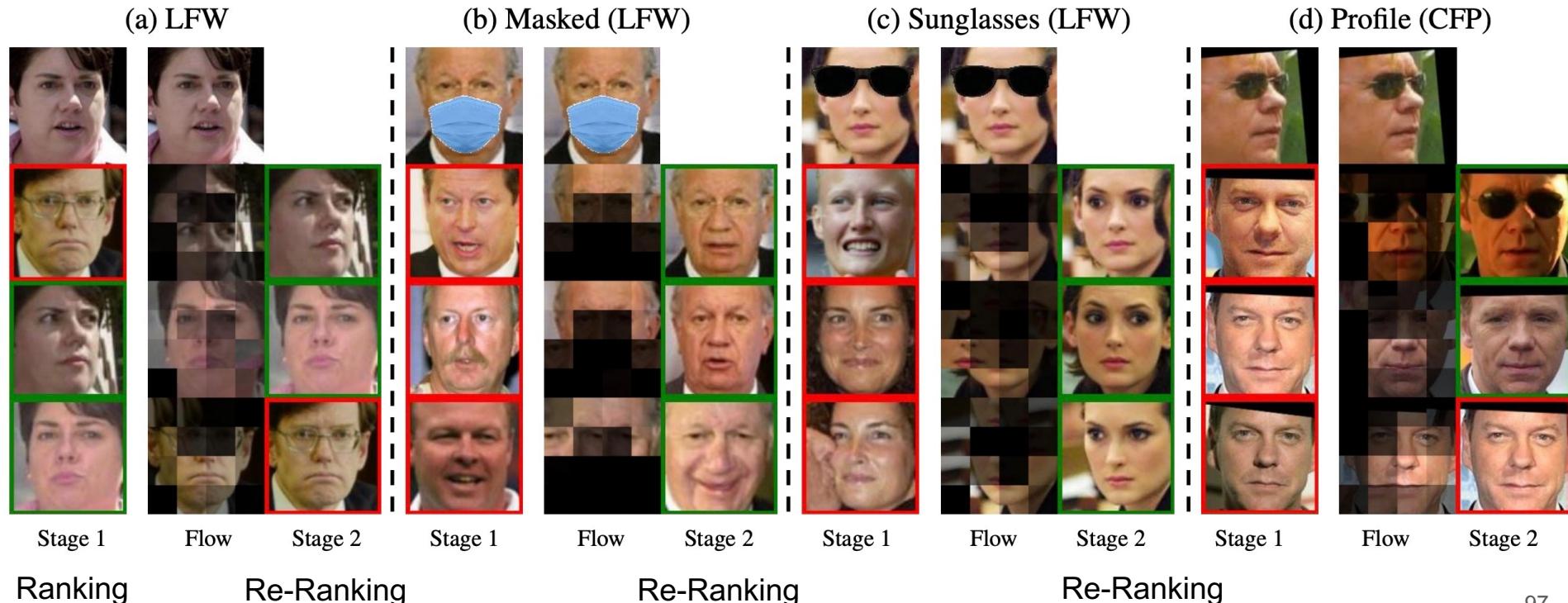


# Cross correlation to filter out irrelevant regions



$$w_{q_i} = \max \left( 0, \langle q_i, \frac{\sum_j^N g_j}{N} \rangle \right), w_{g_j} = \max \left( 0, \langle g_j, \frac{\sum_i^N q_i}{N} \rangle \right)$$

# DeepFace-EMD improves robustness for all tested backbone CNNs



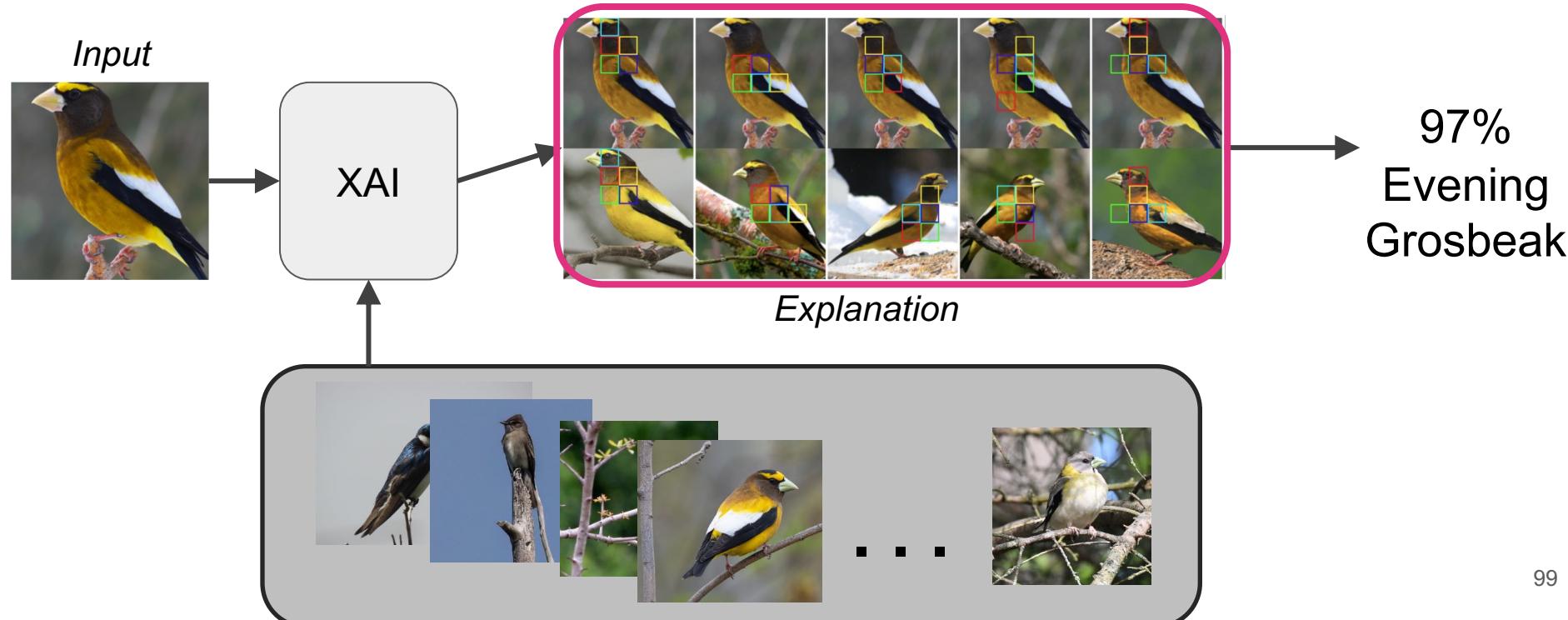
# DeepFace-EMD improves robustness for all tested backbone CNNs

Dataset	Model	Method	P@1	RP	M@R
CALFW (Mask)	ArcFace	ST1	96.81	53.13	51.70
		Ours	<b>99.92</b>	<b>57.27</b>	<b>56.33</b>
	CosFace	ST1	98.54	43.46	41.20
		Ours	<b>99.96</b>	<b>59.85</b>	<b>58.87</b>
	FaceNet	ST1	77.63	39.74	36.93
		Ours	<b>96.67</b>	<b>45.87</b>	<b>44.53</b>
CALFW (Sunglass)	ArcFace	ST1	51.11	29.38	26.73
		Ours	<b>54.95</b>	<b>30.66</b>	<b>27.74</b>
	CosFace	ST1	45.20	25.93	22.78
		Ours	<b>49.67</b>	<b>26.98</b>	<b>24.12</b>
	FaceNet	ST1	21.68	13.70	10.89
		Ours	<b>25.07</b>	<b>15.04</b>	<b>12.16</b>
CALFW (Crop)	ArcFace	ST1	79.13	43.46	41.20
		Ours	<b>92.57</b>	<b>47.17</b>	<b>45.68</b>
	CosFace	ST1	10.99	6.45	5.43
		Ours	<b>25.99</b>	<b>12.35</b>	<b>11.13</b>
	FaceNet	ST1	79.47	44.40	41.99
		Ours	<b>85.71</b>	<b>45.91</b>	<b>43.83</b>

Dataset	Model	Method	P@1	RP	M@R
AgeDB (Mask)	ArcFace	ST1	96.15	39.22	30.41
		Ours	<b>99.84</b>	39.22	<b>33.18</b>
	CosFace	ST1	98.31	38.17	31.57
		Ours	<b>99.95</b>	<b>39.70</b>	<b>33.68</b>
	FaceNet	ST1	75.99	22.28	14.95
		Ours	<b>96.53</b>	<b>24.25</b>	<b>17.49</b>
AgeDB (Sunglass)	ArcFace	ST1	84.64	51.16	44.99
		Ours	<b>87.06</b>	50.40	44.27
	CosFace	ST1	68.93	34.90	27.30
		Ours	<b>75.97</b>	<b>35.54</b>	<b>28.12</b>
	FaceNet	ST1	56.77	27.92	20.00
		Ours	<b>61.21</b>	<b>28.98</b>	<b>21.11</b>
AgeDB (Crop)	ArcFace	ST1	79.92	32.66	26.19
		Ours	<b>92.92</b>	<b>32.93</b>	<b>26.60</b>
	CosFace	ST1	10.11	4.23	2.18
		Ours	<b>19.58</b>	<b>4.95</b>	<b>2.76</b>
	FaceNet	ST1	80.80	31.50	24.27
		Ours	<b>86.74</b>	<b>31.51</b>	<b>24.32</b>

# Solution: “Slow-thinking” System 2, Explainable AIs

- First explain, Then decide
- Harness external knowledge-bases to make informed decisions

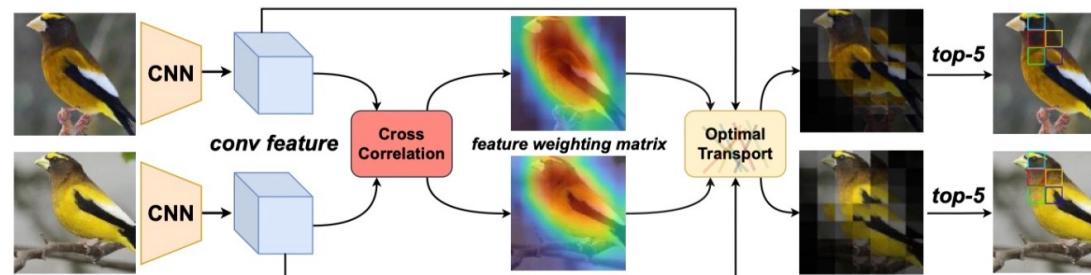




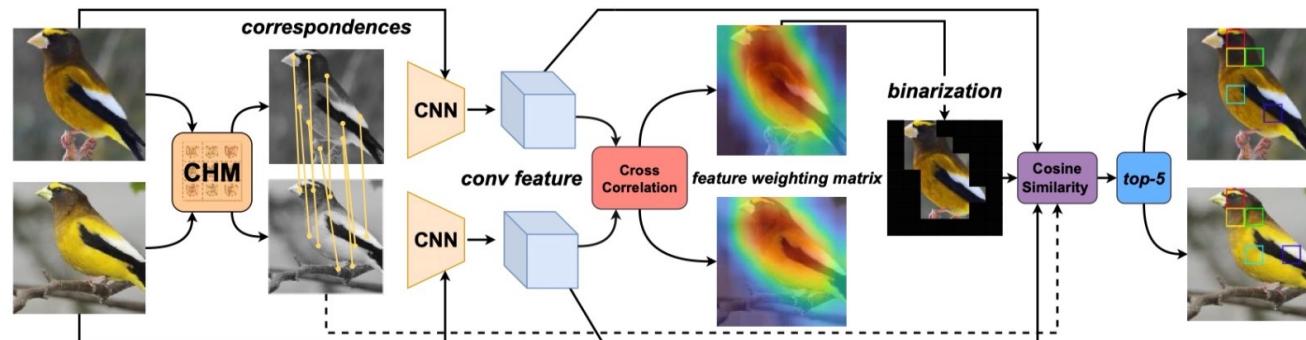
Query Image

**Idea:** Rank using image similarity, then, re-rank using patch-wise similarity

1. Zoom-in to find important, discriminative patches
2. Use only 5 most important patches per image to make (zero-shot) classification decisions.



(a) EMD-Corr: First compute patch-wise similarity, and then find correspondences via solving EMD [29, 75].



(b) CHM-Corr: First find correspondences via CHM [49], and then compute patch-wise similarity.

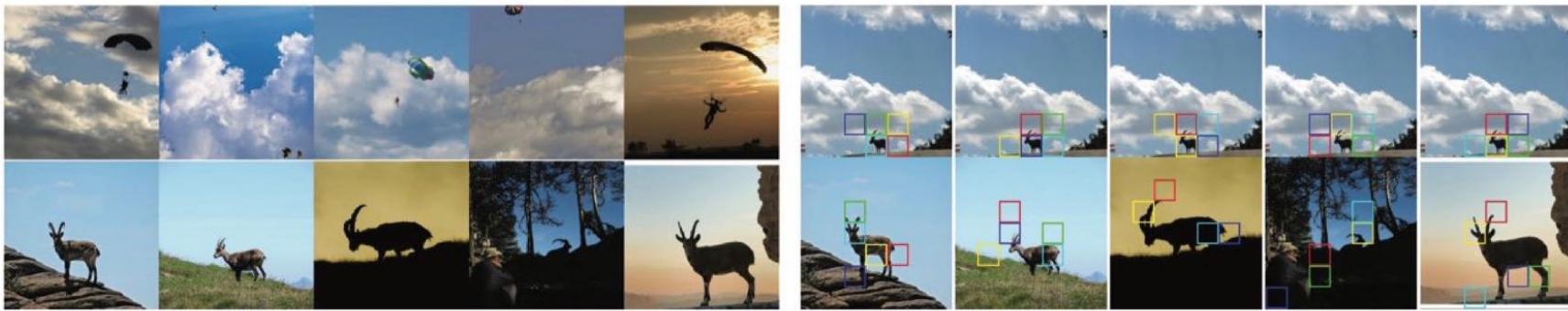
# Zoom-in to find important, discriminative patches

Test set	Features (a)	Training set (b)	ResNet-50	kNN	EMD-Corr	CHM-Corr	CHM-Corr+
ImageNet [61]	ImageNet	ImageNet	<b>76.13</b>	74.77	74.93 (-1.20)	74.40 (-1.73)	n/a
ImageNet-ReaL [14]	ImageNet	ImageNet	<b>83.04</b>	82.05	82.32 (-0.72)	81.97 (-1.07)	n/a
ImageNet-R [35]	ImageNet	ImageNet	36.17	36.18	<b>37.75 (+1.58)</b>	37.62 (+1.45)	n/a
ImageNet Sketch [70]	ImageNet	ImageNet	24.09	24.72	25.36 (+1.27)	<b>25.61 (+1.52)</b>	n/a
DAMageNet [18]	ImageNet	ImageNet	5.93	7.59	<b>8.16 (+2.23)</b>	8.10 (+2.17)	n/a
Adversarial Patch [15]	ImageNet	ImageNet	55.04	59.30	59.43 (+4.39)	<b>59.86 (+4.82)</b>	n/a
CUB [69]	ImageNet	CUB	n/a	54.72	<b>60.29</b>	53.65	49.63
CUB [69]	iNaturalist [68]	CUB	<b>85.83</b>	85.46	84.98 (-0.85)	83.27 (-2.56)	81.54

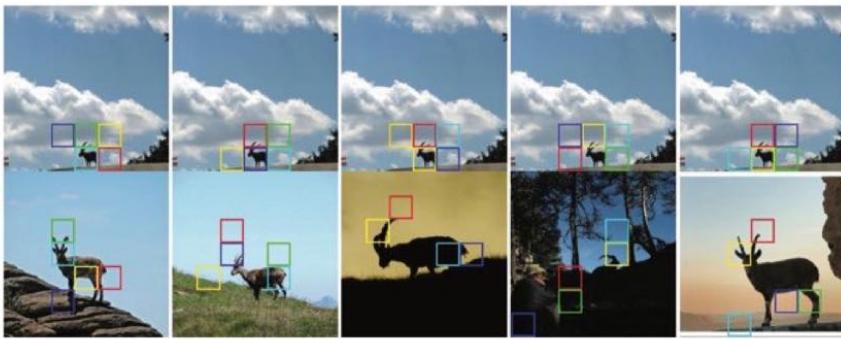
# Zoom-in to find important, discriminative patches



groundtruth:  
ibex



(a) Explanations for kNN's **parachute** decision (top) and CHM-NN (bottom)



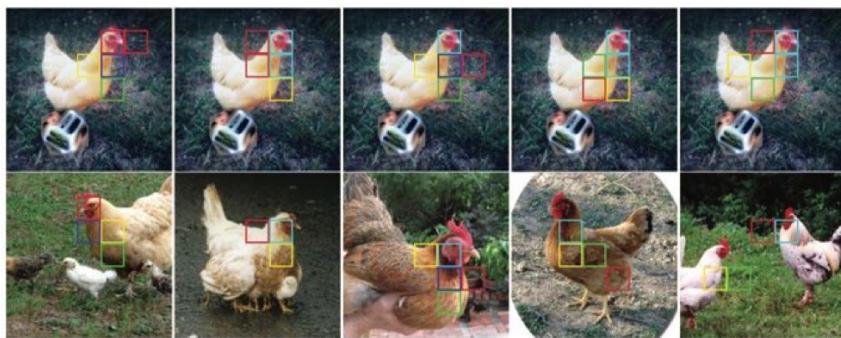
(b) Explanations for CHM-Corr's **ibex** decision



groundtruth:  
hen



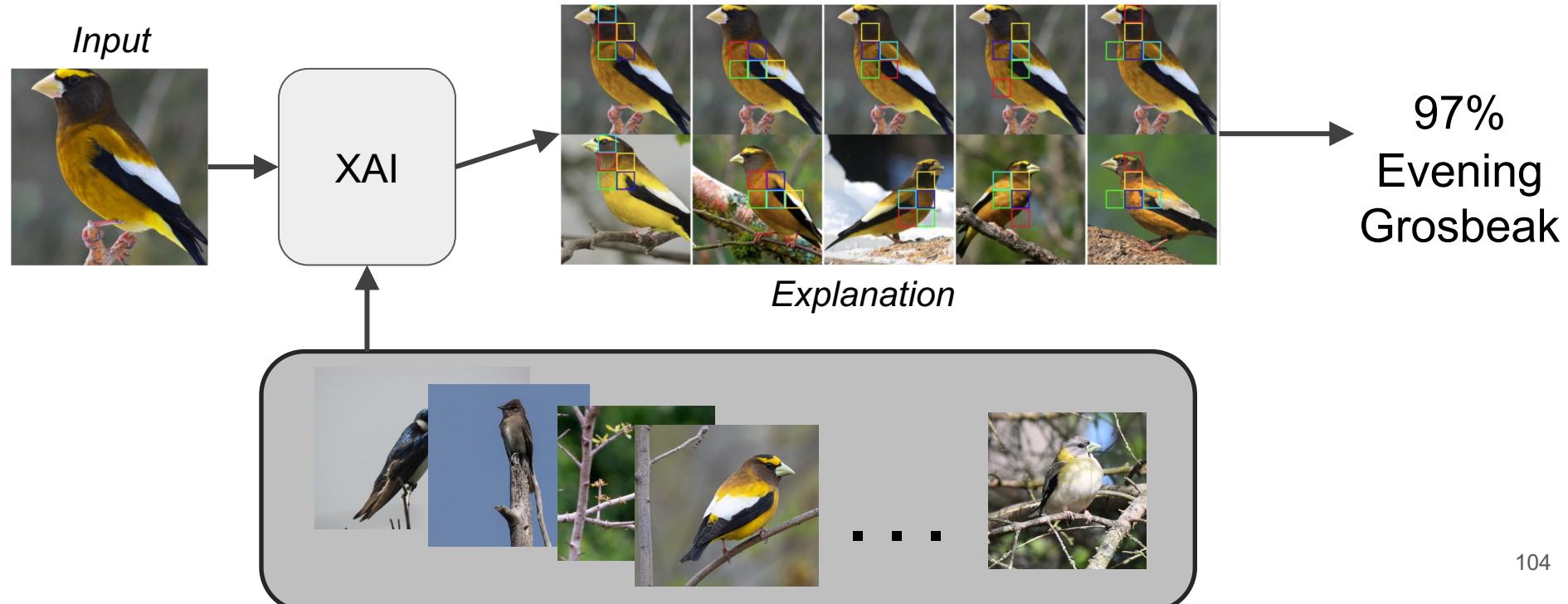
(a) Explanations for kNN's **toaster** decision (top) and EMD-NN (bottom)



(b) Explanations for EMD-Corr's **hen** decision

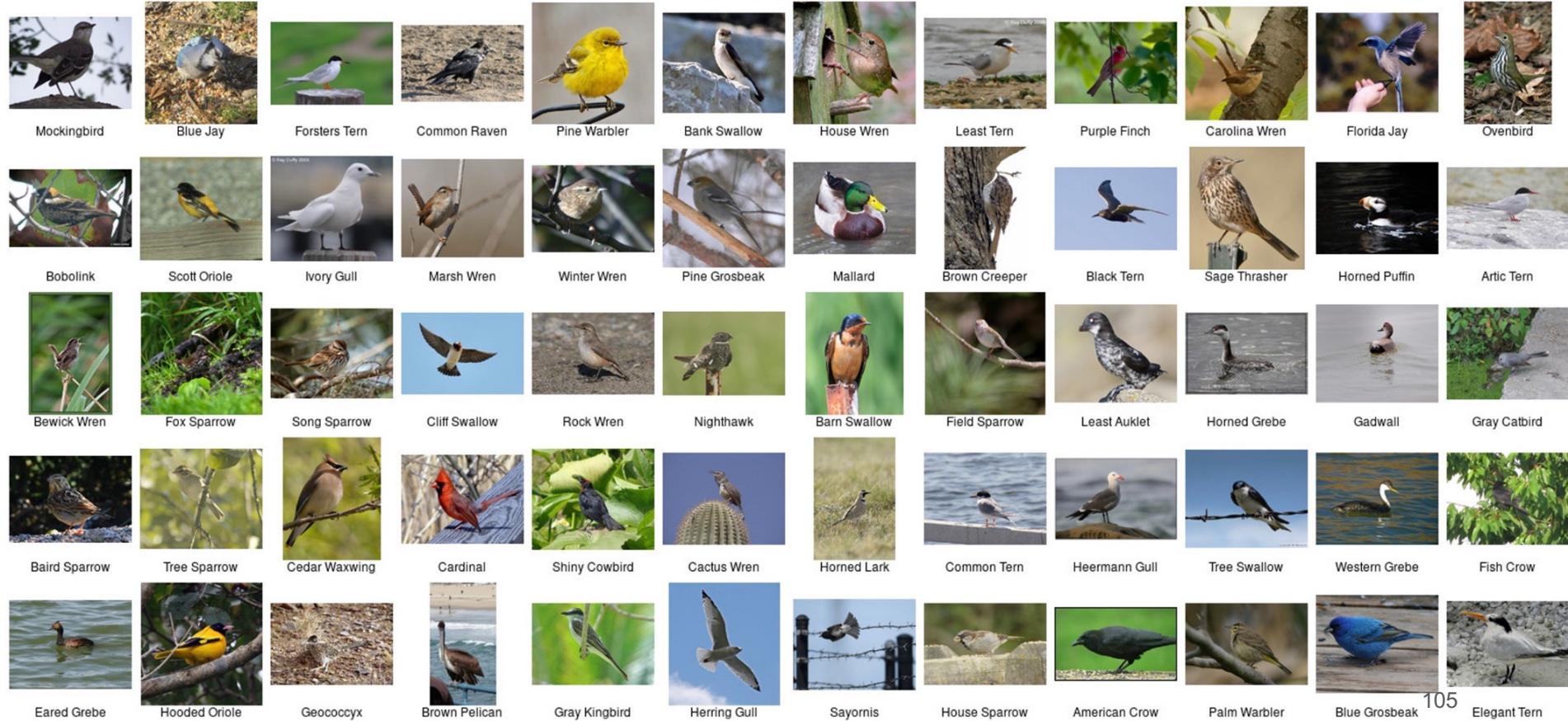
# Solution: “Slow-thinking” System 2, Explainable AIs

- First explain, Then decide
- Harness external knowledge-bases to make informed decisions



# Fine-grained Bird identification (CUB200-2011) 200 classes

200 bird species



# Fine-grained Bird identification (CUB200-2011) 200 classes

***Task:*** Choose one among 200 classes



Acadian Flycatcher

American Crow

American Goldfinch

...

Common Raven

...

House Sparrow

House Wren

...

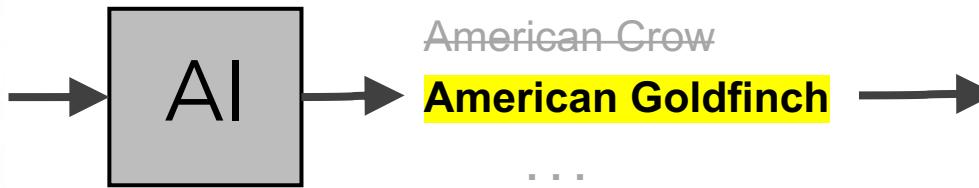


# Fine-grained Bird identification

Yes / No question

~~Task: Choose one among 200 classes~~

I'm 60% confident this is...



Agree  
vs  
Disagree

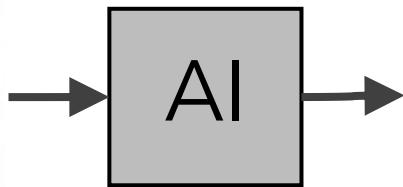


# Fine-grained Bird identification

Yes / No question

~~Task: Choose one among 200 classes~~

I'm 30% confident this is...



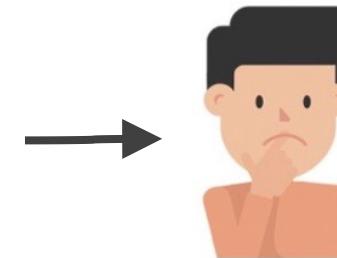
Acadian Flycatcher

American Crow

American Goldfinch

...

**Evening Grosbeak**



Agree

vs

Disagree

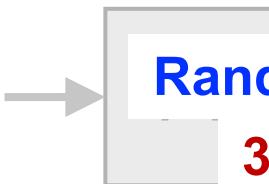


# Nearest neighbors hurts user accuracy

*Task: Choose one among 200 classes*



I'm 30% confident this is



Method	CUB	
	Users	Accuracy
Random explanations	60	65.50 ± 7.46
3-NN explanations	59	64.75 ± 7.14

Evening Grosbeak



## 2. AI top-1 label and confidence score

Sam thinks this is 97% junco



## 1. query image

## 4. user decision

Is this junco?

Yes

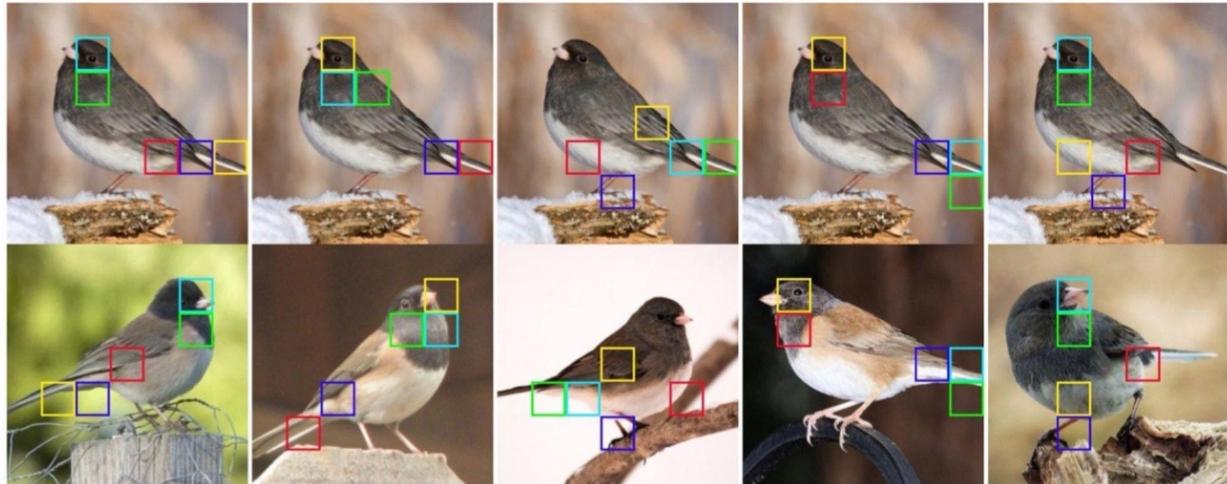
No

## 2. AI top-1 label and confidence score

Sam thinks this is **97% junco** after comparing its content in the 5 colored boxes with the corresponding ones in junco images.



1. query image



junco

## 3. explanation

## 4. user decision

Is this **junko**?

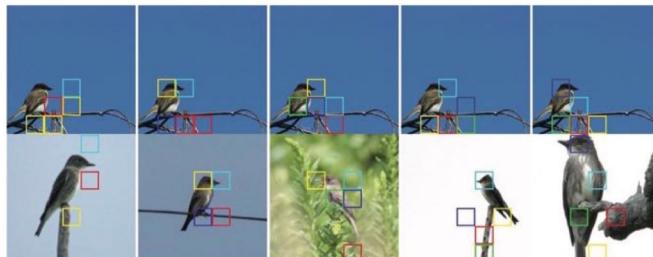
Yes

No

# Visual correspondence improves user accuracy



(a) kNN (top) and CHM-NN (bottom) explanation



(b) CHM-Corr visual correspondence-based explanations

Method	CUB	
	Users	Accuracy
Random explanations )	60	<b><math>65.50 \pm 7.46</math></b>
3-NN explanations	59	$64.75 \pm 7.14$
EMD-NN	59	$64.12 \pm 7.07$
EMD-Corr	58	<b><math>67.64 \pm 7.44</math></b>
CHM-NN	60	$65.72 \pm 8.14$
CHM-Corr	59	<b><math>69.72 \pm 9.08</math></b>

# Long-term Research Goals

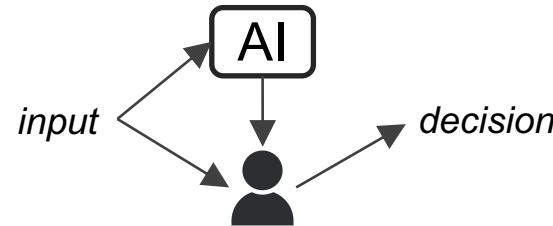
## Trustworthy AI

1. Build AIs that are accurate in edge cases  
(and common cases)

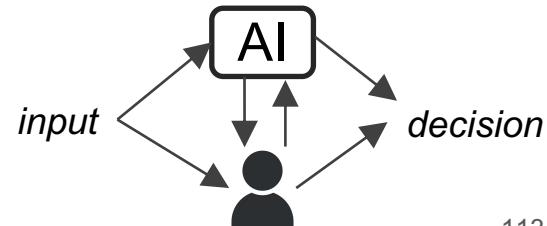


## Explainable AI

2. Build AIs that maximize human-AI team accuracy



3. Build AIs that humans can debug and edit  
(their decision-making process)



# Long-term Research Goals

## Trustworthy AI

1. Build AIs that are accurate in edge cases  
(and common cases)

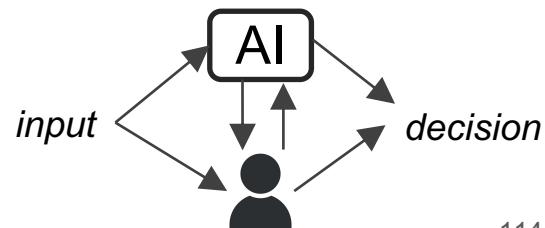


## Explainable AI

2. Build AIs that maximize human-AI team accuracy

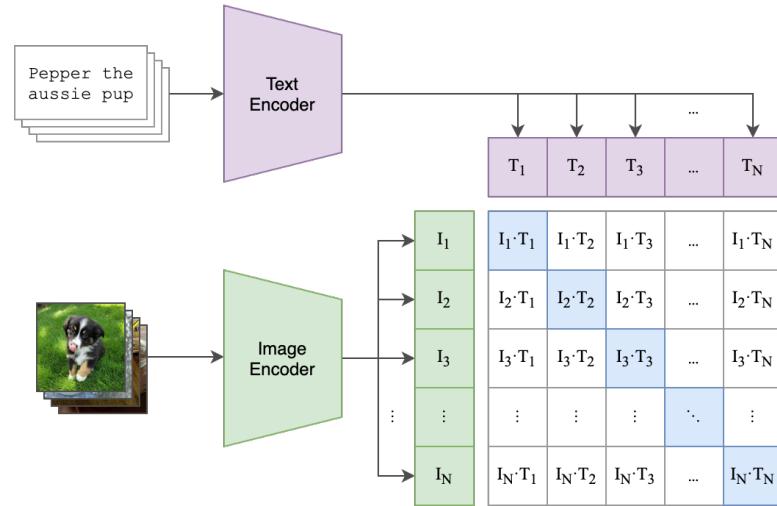
# Part 3

3. Build AIs that humans can debug and edit  
(their decision-making process)

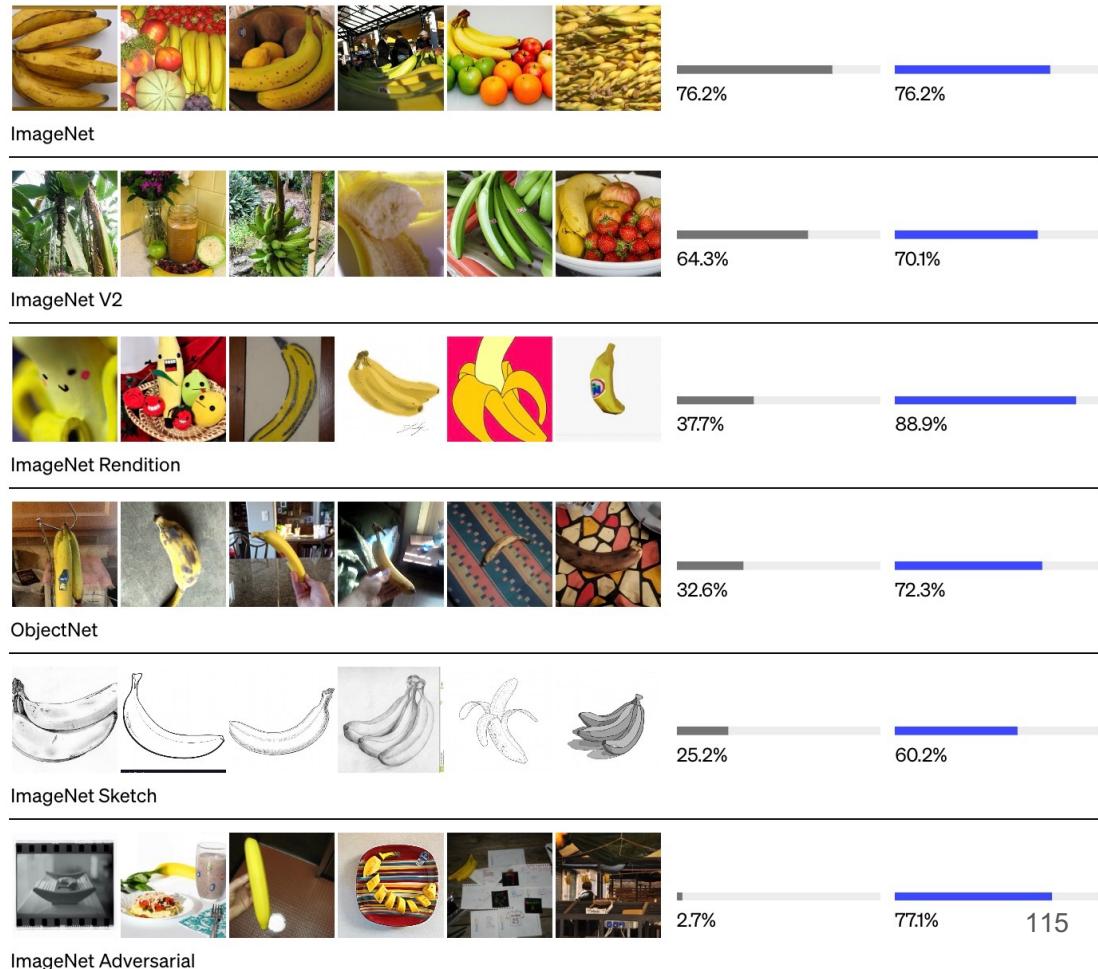




### (1) Contrastive pre-training

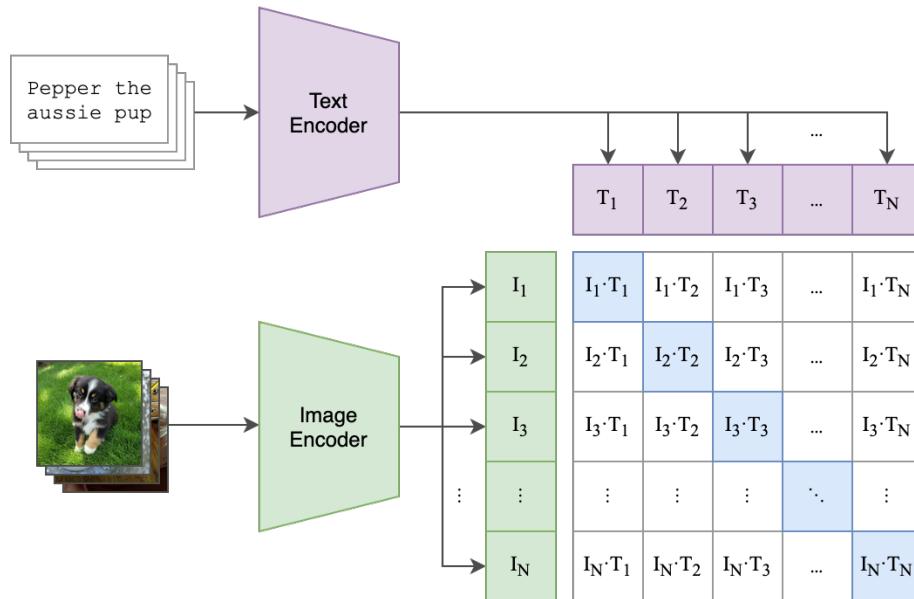


### Dataset

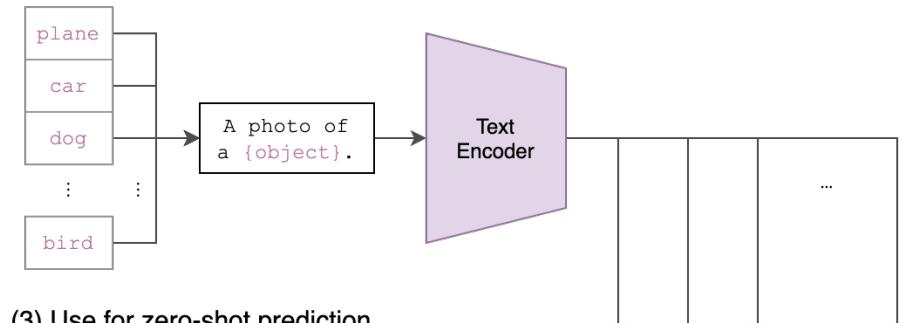


# What is a key limitation of CLIP classifiers?

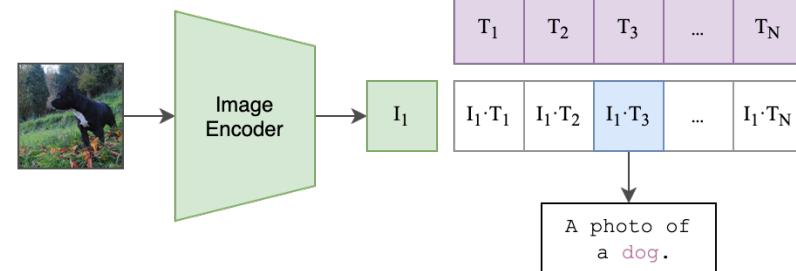
(1) Contrastive pre-training



(2) Create dataset classifier from label text

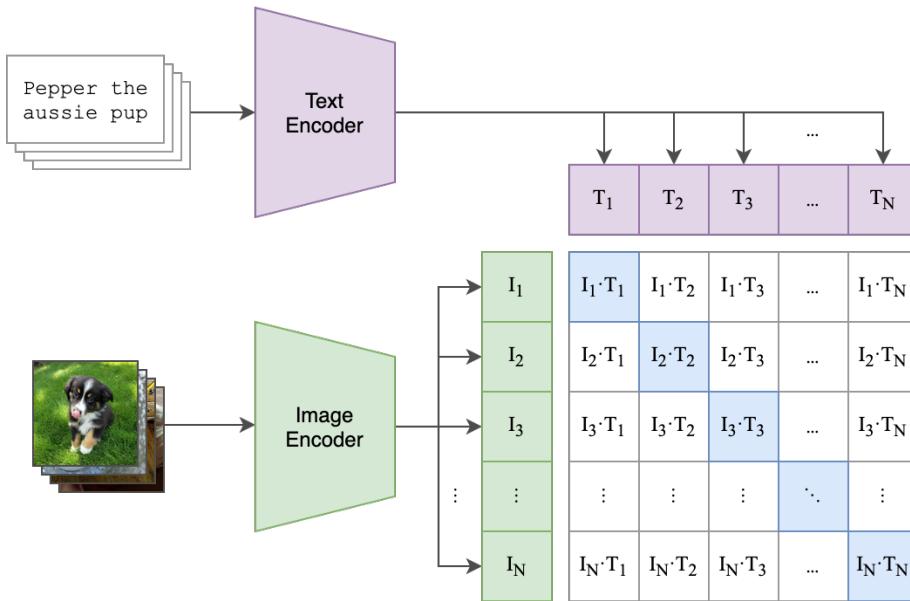


(3) Use for zero-shot prediction

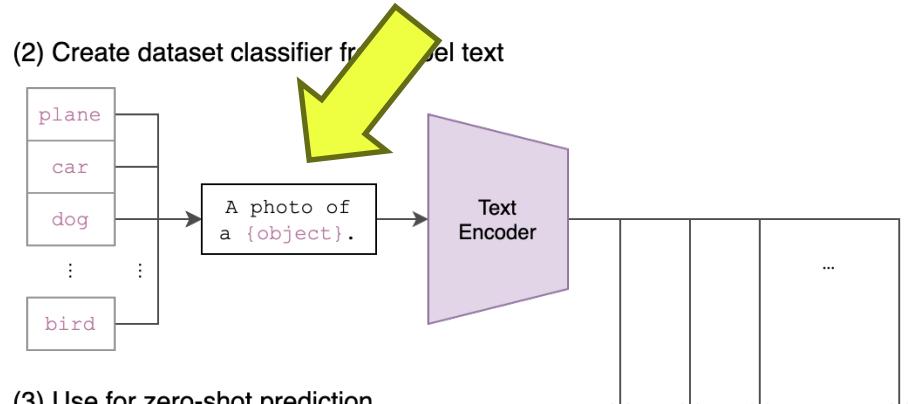


# What is a key limitation of CLIP classifiers?

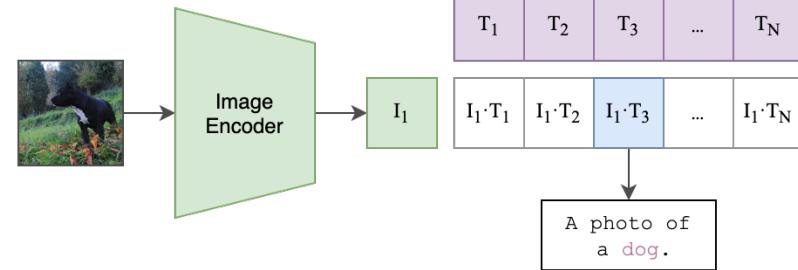
(1) Contrastive pre-training



(2) Create dataset classifier from novel text



(3) Use for zero-shot prediction

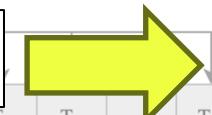




What is a key limitation of CLIP classifiers?

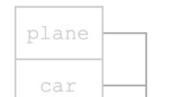
(1)

A photo of a **Blue Jay**



A photo of a **Cyanocitta cristata**

(2) Create dataset classifier from label text



Methods

CUB CUB<sub>sci</sub> NABirds NABirds<sub>sci</sub> iNaturalist iNaturalist<sub>sci</sub>

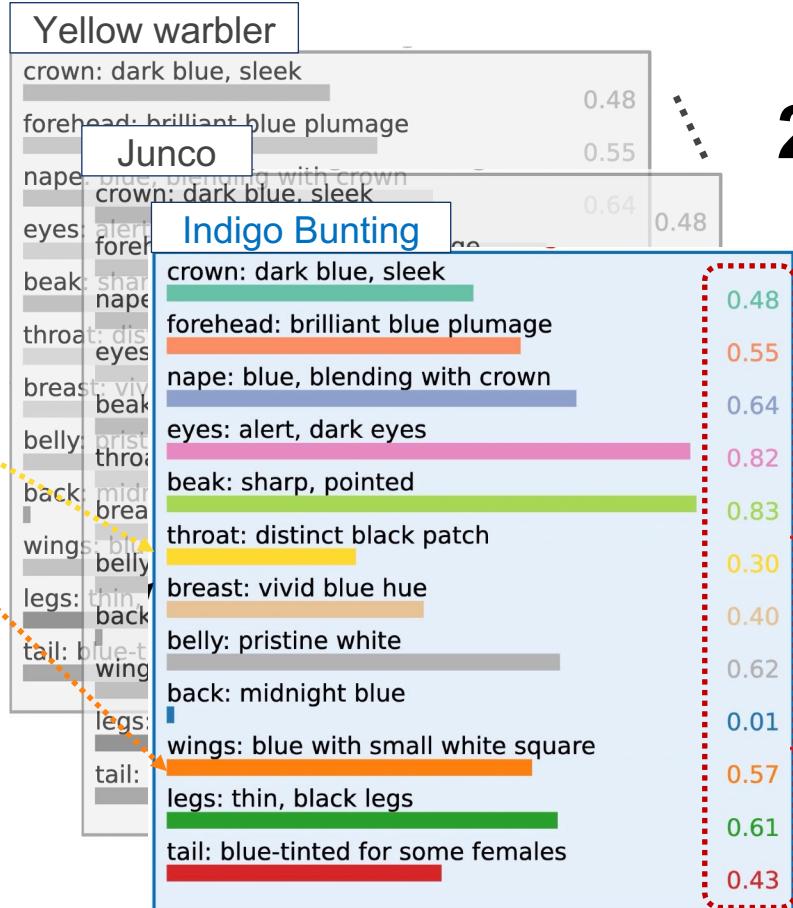
CLIP (Radford et al., 2021) 51.95 5.95 39.35 4.73 16.36 2.03

M&V (Menon & Vondrick, 2022) 53.78 7.66 41.01 6.27 17.57 2.87

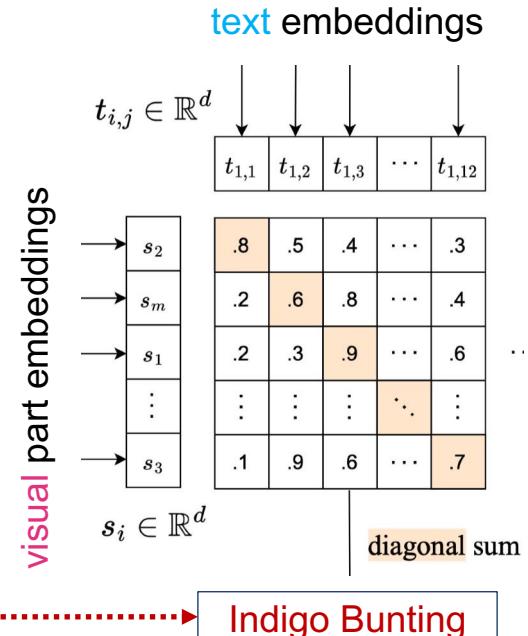
PEEB (ours) **64.33** **69.03** **25.74**

# Part-based, Explainable, Editable Language Bottleneck

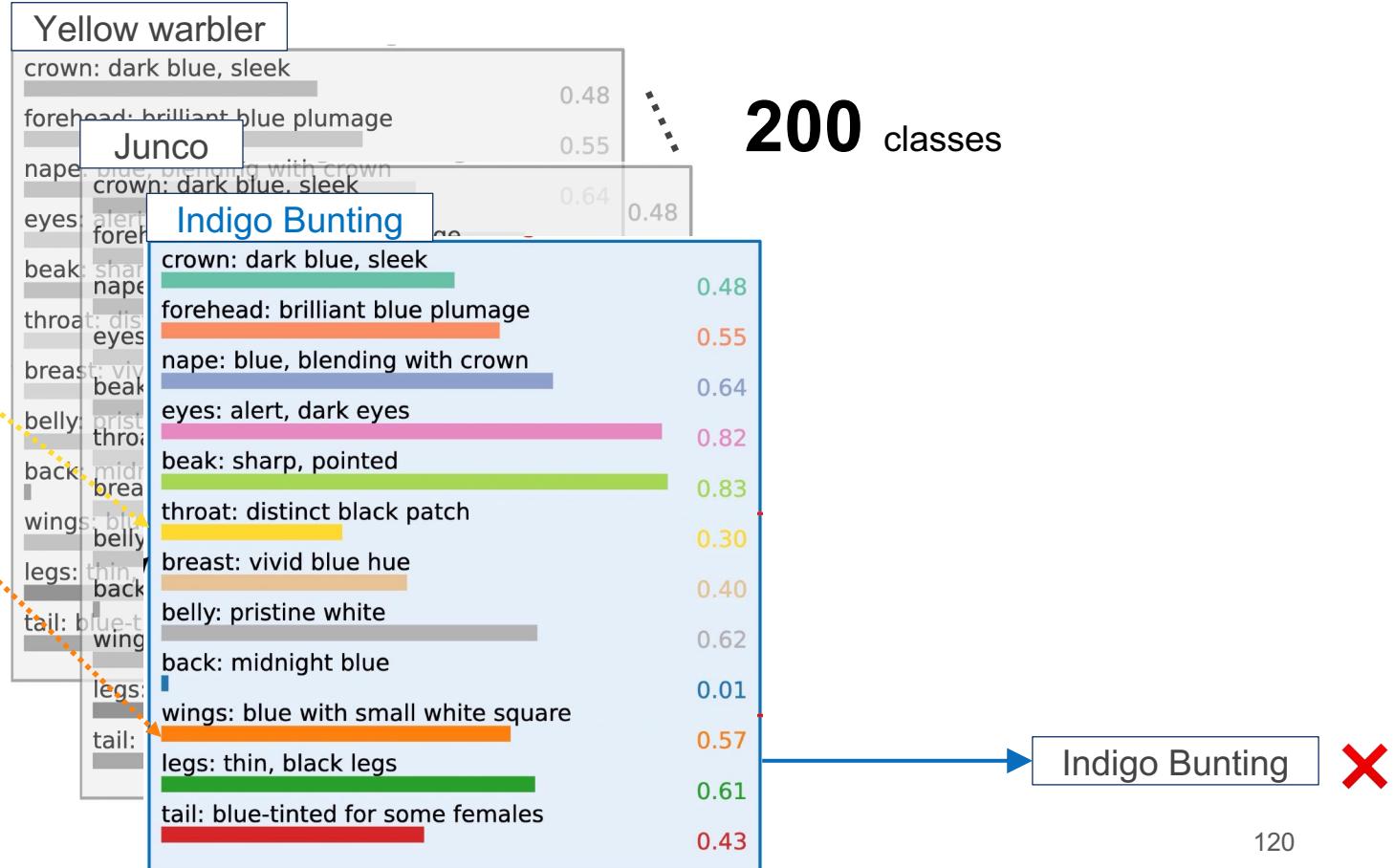
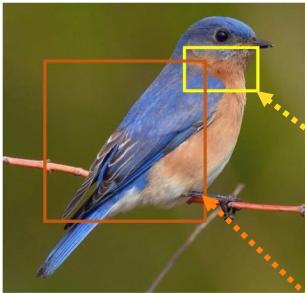
NAACL 24



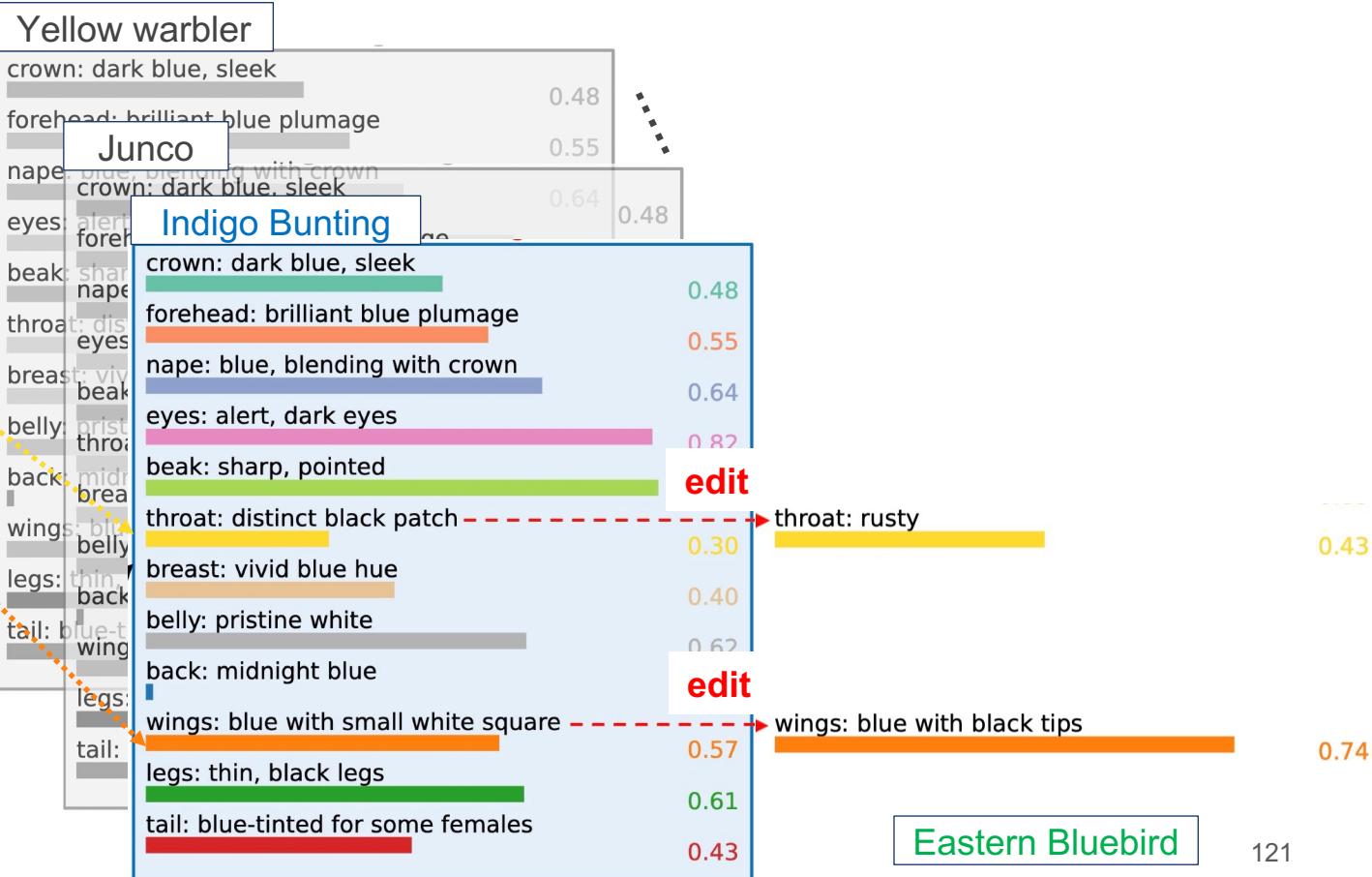
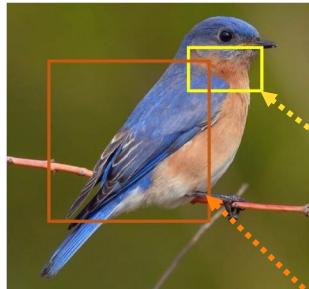
**200** classes



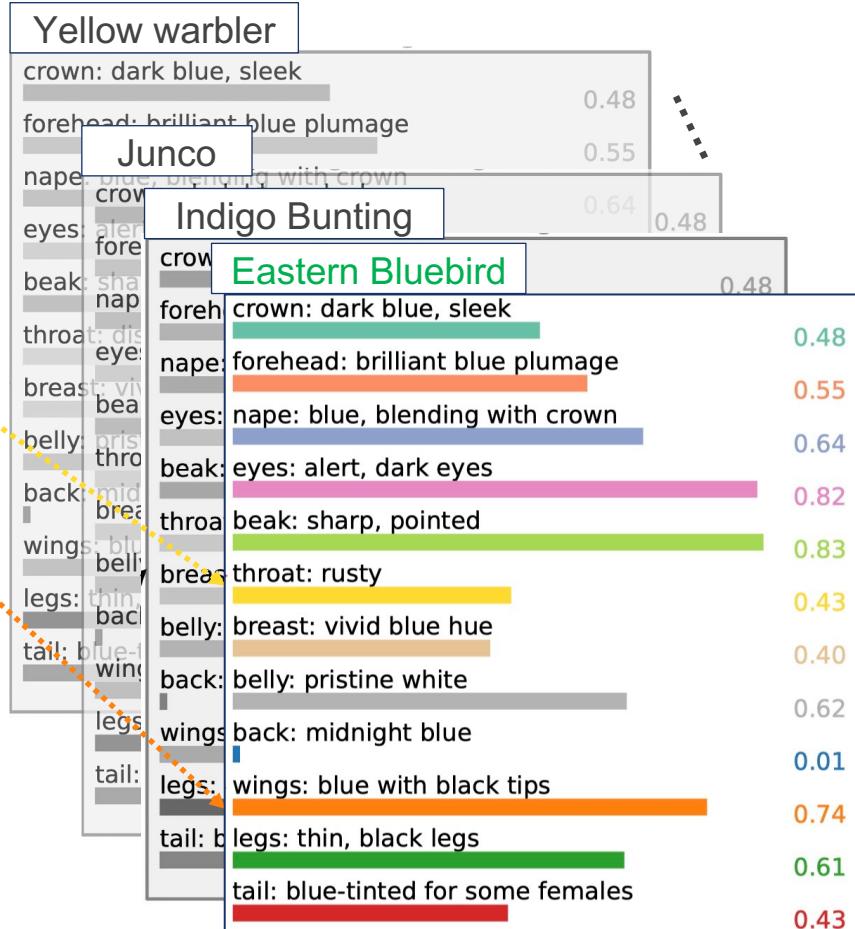
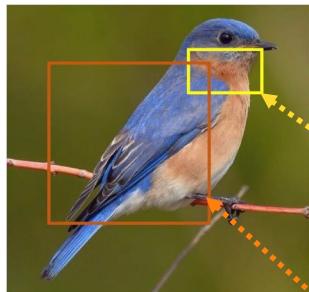
Part-based, Explainable, Editable Language Bottleneck NACCL 24



# Part-based, Explainable, Editable Language Bottleneck



# Part-based, Explainable, Editable Language Bottleneck



200 + 1 classes

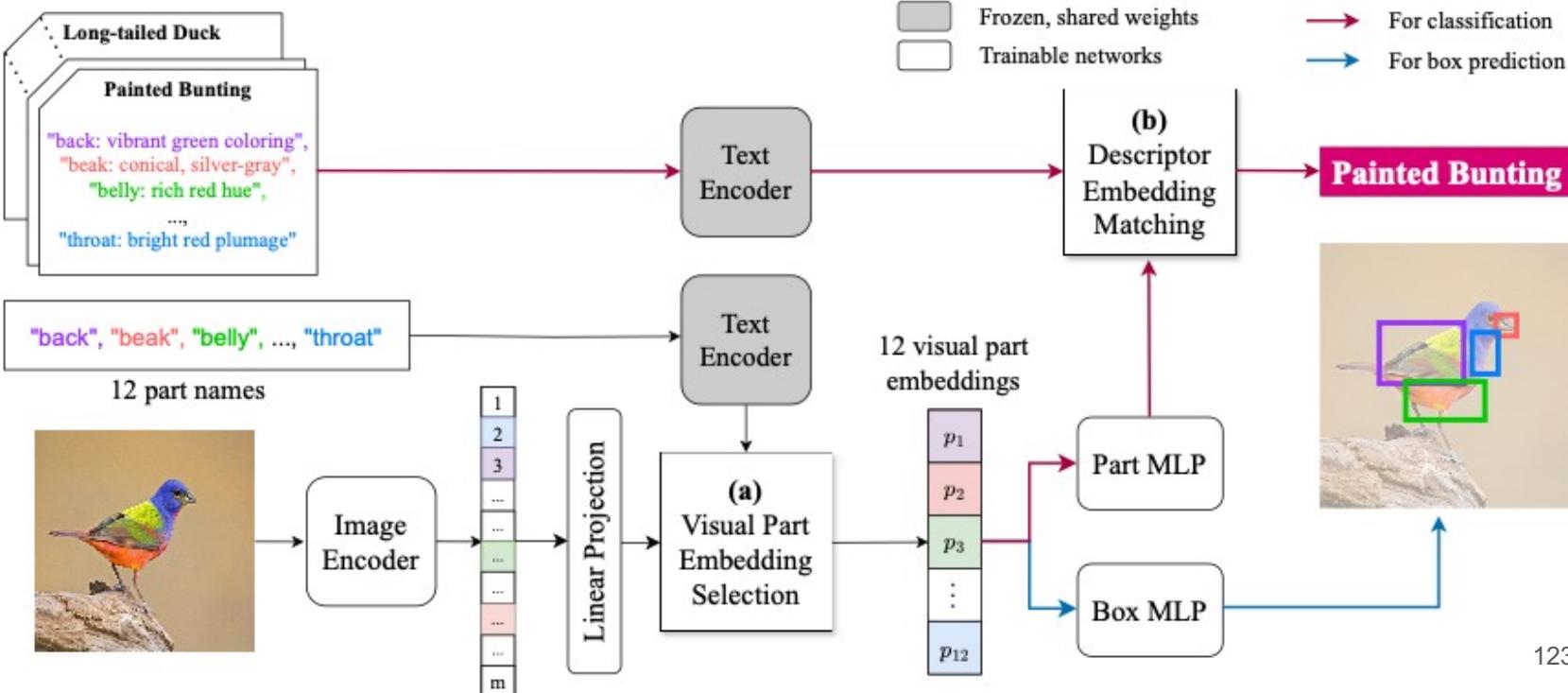
Eastern Bluebird



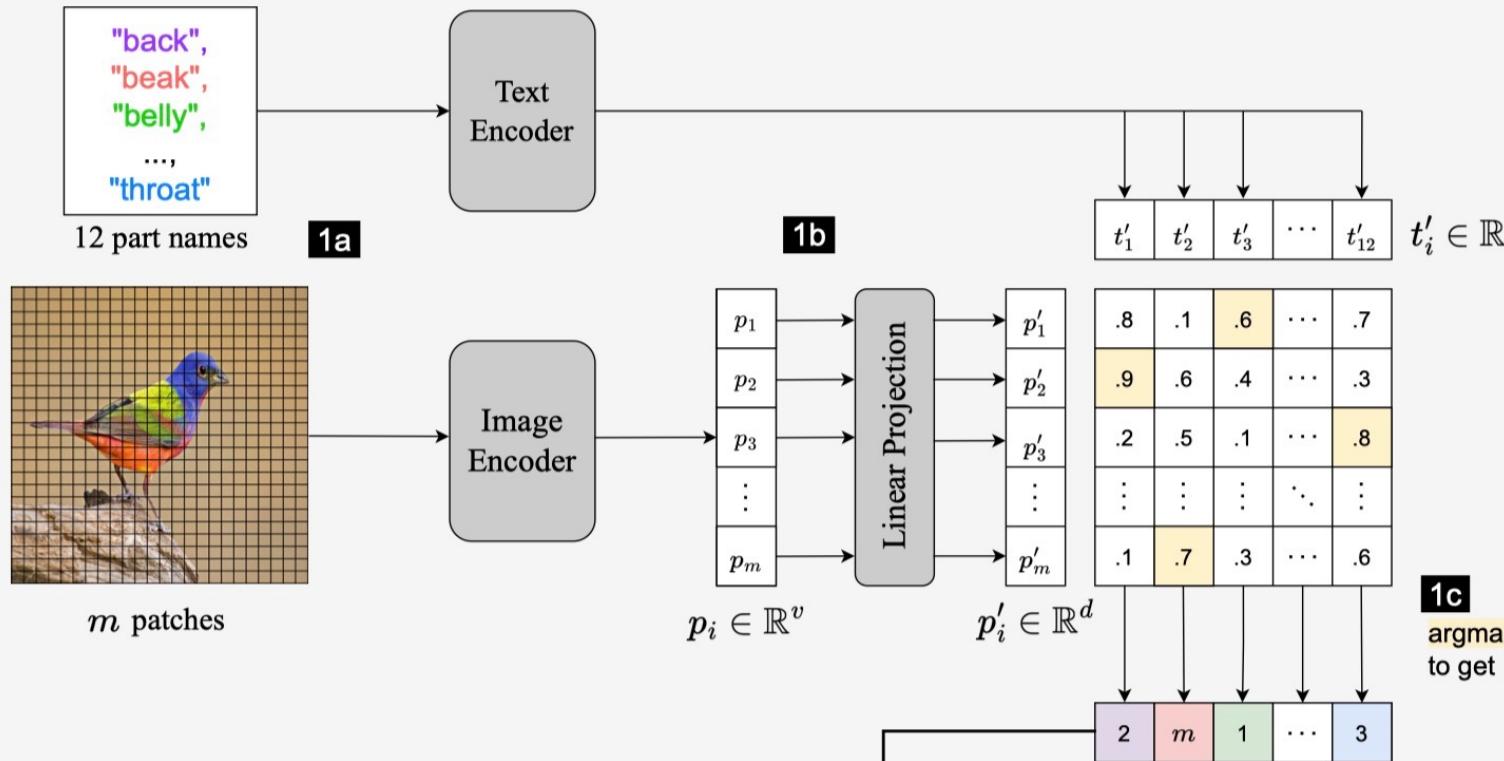
Eastern Bluebird

# Part-based, Explainable, Editable Language Bottleneck

- Build Bird-11K, the largest bird-image dataset ever. **290,000** images and **11,000** images  
~ *all birds on Earth*
- Collect an additional 55,000 (10,534 classes) from Macaulay Library
- Use GPT-4 to generate descriptors
- Contrastive learning + finetuning on a target dataset

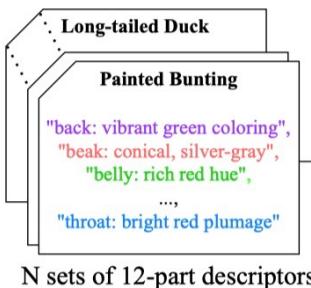


## Step 1: Part embeddings selection from OWL-ViT



## Step 2: Contrastive training between parts and descriptions

## Step 2: Contrastive training between parts and descriptions



Text Encoder

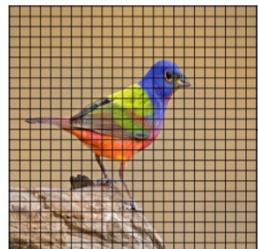
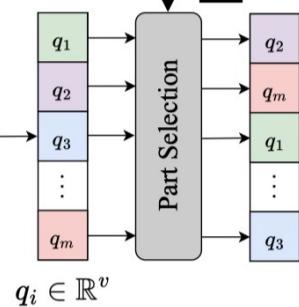


Image Encoder

2a



input for Part Selection in step 2

Part Selection

Part MLP

2b

2c

$t_{i,j} \in \mathbb{R}^d$

$s_2$	.8	.5	.4	...	.3
$s_m$	.2	.6	.8	...	.4
$s_1$	.2	.3	.9	...	.6
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$s_3$	.1	.9	.6	...	.7

$s_i \in \mathbb{R}^d$

Class 1: Painted Bunting

Class N: Long-tailed Duck

## Pre-training Step 1

$t_{N,1}$	.9	.2	.3	...	.1
$t_{N,2}$	.2	.7	.6	...	.3
$t_{N,3}$	.5	.4	.8	...	.6
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$t_{N,12}$	.1	.8	.2	...	.5

$t_{N,1}$	.9	.2	.3	...	.1
$t_{N,2}$	.2	.7	.6	...	.3
$t_{N,3}$	.5	.4	.8	...	.6
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$t_{N,12}$	.1	.8	.2	...	.5

2d

1	0	0	...	0
0	1	0	...	0
0	0	1	...	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
0	0	0	...	1

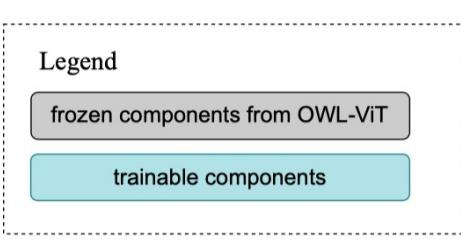
0	0	0	...	0
0	0	0	...	0
0	0	0	...	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
0	0	0	...	125

# Loss 1

1	0	0	...	0
0	1	0	...	0
0	0	1	...	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
0	0	0	...	1

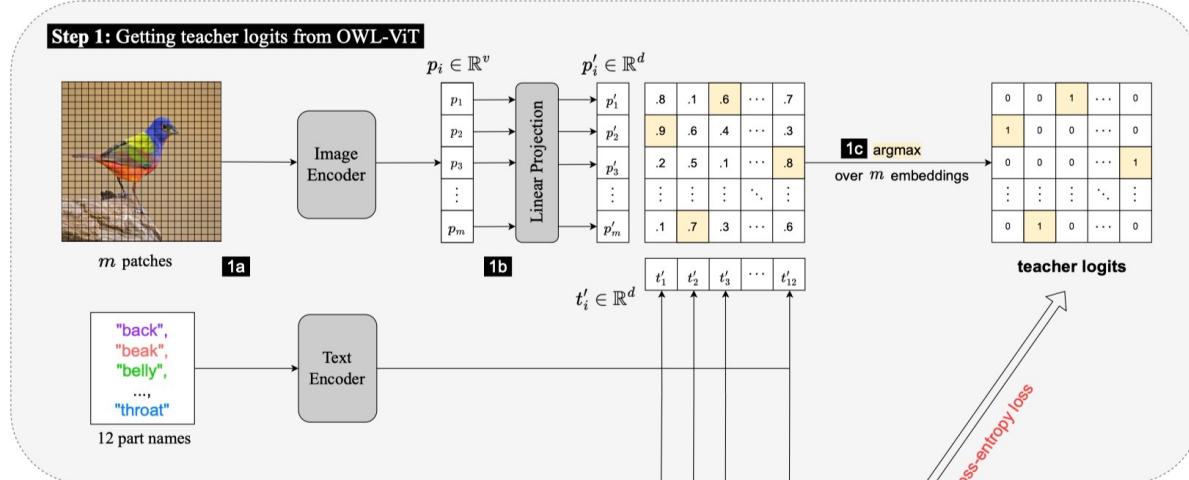
$t_{N,1}$	.9	.2	.3	...	.1
$t_{N,2}$	.2	.7	.6	...	.3
$t_{N,3}$	.5	.4	.8	...	.6
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$t_{N,12}$	.1	.8	.2	...	.5

0	0	0	...	0
0	0	0	...	0
0	0	0	...	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
0	0	0	...	125



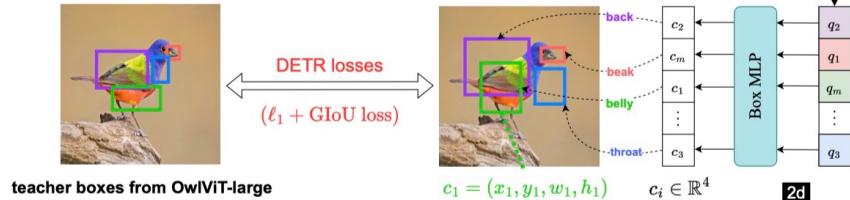
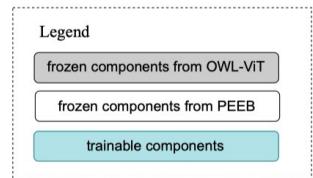
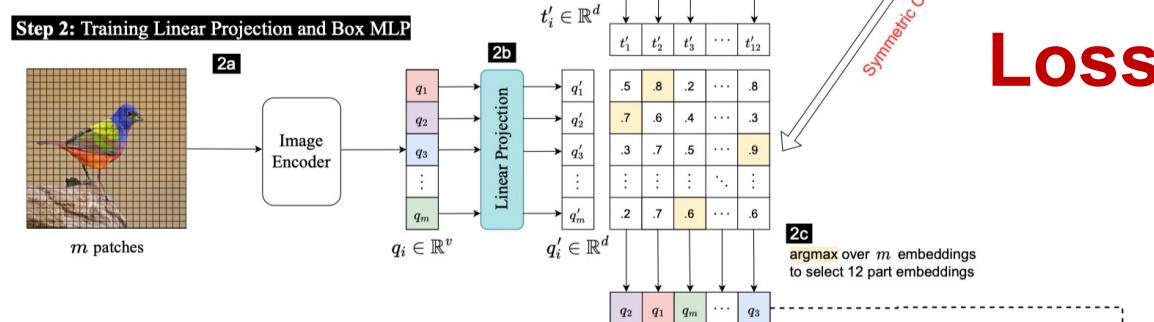
# Pre-training Step 2

- Loss 1: SCE (part name)
- Loss 2: SCE (part descriptor)
- Loss 3: DETR



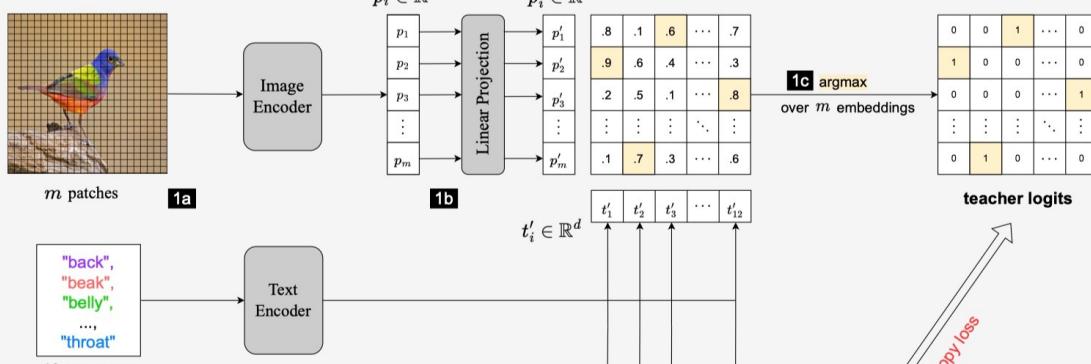
## Loss 2

**Step 2: Training Linear Projection and Box MLP**

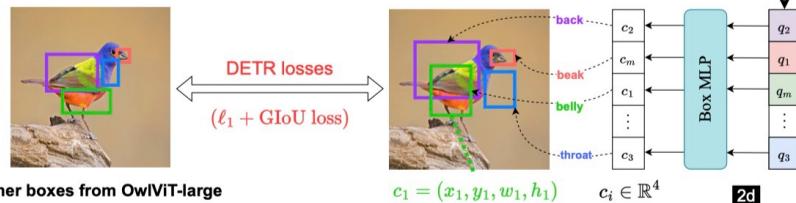
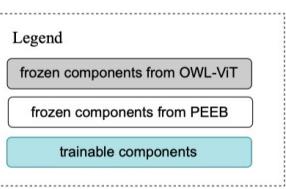
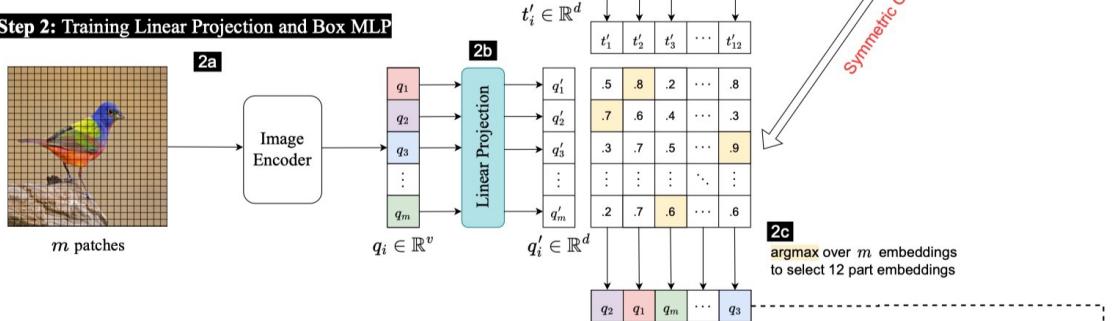


## Loss 3

### Step 1: Getting teacher logits from OWL-ViT



### Step 2: Training Linear Projection and Box MLP



## Finetuning on CUB-200

- Loss 1: SCE (part name)
- Loss 2: CE (part descriptor)
- Loss 3: DETR

# PEEB obtains SotA zero-shot performance on birds

CLIP depends heavily on the known class names in the prompt “*a photo of a painted bunting*”

Methods	CUB	$CUB_{sci}$	NABirds	$NABirds_{sci}$	iNaturalist	$iNaturalist_{sci}$
CLIP (Radford et al., 2021)	51.95	5.95	39.35	4.73	16.36	2.03
M&V (Menon & Vondrick, 2022)	53.78	7.66	41.01	6.27	17.57	2.87
PEEB (ours)		<b>64.33</b>		<b>69.03</b>		<b>25.74</b>

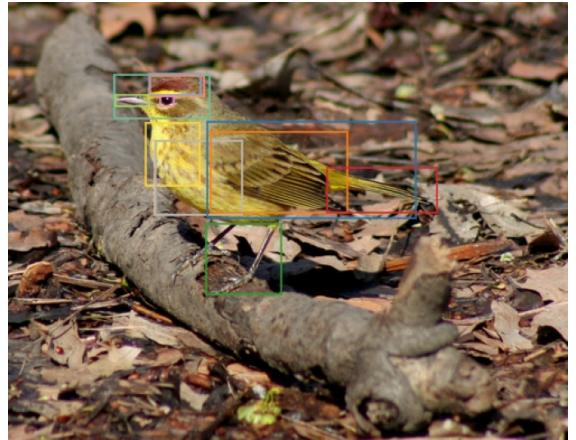
Table 2: In a CLIP’s zero-shot setting (CZSL), our method’s top-1 accuracy is +8 to +28 points higher than the two baselines. When using novel class names (or scientific names which are less common), our method is around 10× better than the others.

CLIP’s zero-shot: CZSL (test on both seen and unseen classes,  
BUT samples of unseen classes may be in the training set)

# PEEB obtains SotA **generalized** zero-shot performance

Method	Methods	CUB			NABirds		
		Seen	Unseen	Harmonic	Seen	Unseen	Harmonic
CLORE <sub>CLIP</sub>	(Han et al., 2022)	65.80	39.10	49.05			
PEEB (ours)		<b>80.78</b>	<b>41.74</b>	<b>55.04</b>		n/a	
PEEB		SCS	SCE	Harmonic	SCS	SCE	Harmonic
		(Easy)	(Hard)		(Easy)	(Hard)	
S <sup>2</sup> GA-DET	(Ji et al., 2018)	42.90	10.90	17.38	<b>39.40</b>	9.70	15.56
GRZSL	(Zhu et al., 2018)	44.08	14.46	21.77	36.36	9.04	14.48
DGRZSL	(Kousha & Brubaker, 2021)	<b>45.48</b>	14.29	21.75	37.62	8.91	14.41
PEEB (ours)		44.66	<b>20.31</b>	<b>27.92</b>	28.26	<b>24.34</b>	<b>26.15</b>

Table 3: Over all 5 splits, our model consistently outperforms other methods (top-1 accuracy). Specifically, PEEB outperforms other methods in the hard split (SCE) by (+6 to +15) points, highlighting its generalization capability.

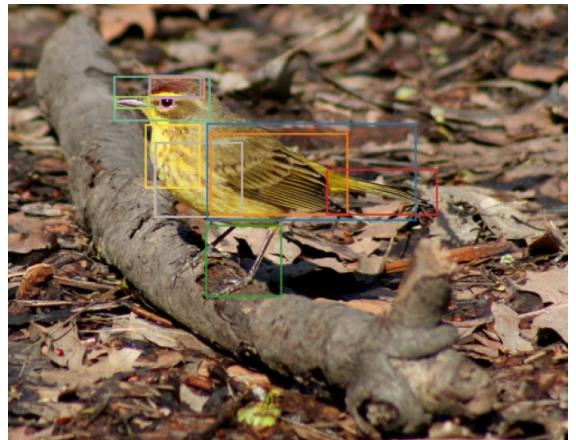


Our prediction: **palm warbler** 0.819  
because of the following...

crown: orange-yellow with pale edges  
forehead: yellowish with faint markings  
nape: olive-brown, blending into the back  
eyes: small and dark, framed by eye-ring  
beak: short and sharp, black-colored  
throat: bright yellow, blending into the breast  
breast: bright yellow with dark streaks  
belly: creamy white with faint streaks  
back: olive-brown back with streaks  
wings: olive-brown with white-edged feathers  
legs: long and skinny, with blackish coloring  
tail: short and dark, with white outer feathers

M&V's prediction: **prairie warbler** 0.002  
but we rejected it because...

0.696	crown: yellowish-green	0.000
0.688	forehead: yellow with black markings	0.309
0.722	nape: greenish-yellow	0.000
0.483	eyes: dark with thin white eye-ring	0.212
0.475	beak: small and pointed	0.149
0.672	throat: bright yellow	0.173
0.614	breast: bright yellow with faint streaks	0.551
0.624	belly: yellowish with light brown streaks	0.306
0.688	back: olive-green with faint streaks	0.100
0.575	wings: dark grayish-brown with white streaks	0.220
0.645	legs: pinkish-brown	0.000
0.699	tail: dark grayish-brown with white edges	0.142

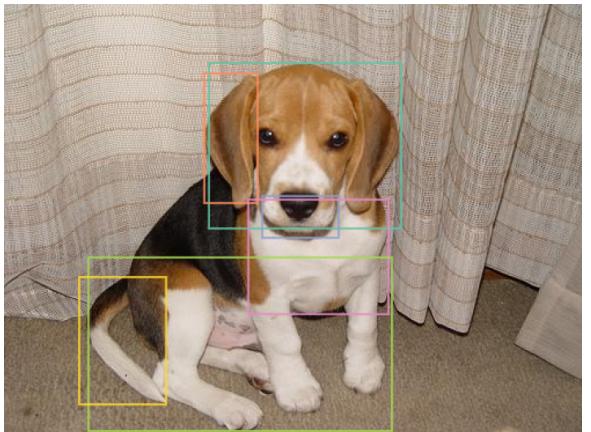


Our prediction: **palm warbler** 0.819  
because of the following...

crown: orange-yellow with pale edges  
forehead: yellowish with faint markings  
nape: olive-brown, blending into the back  
eyes: small and dark, framed by eye-ring  
beak: short and sharp, black-colored  
throat: bright yellow, blending into the breast  
breast: bright yellow with dark streaks  
belly: creamy white with faint streaks  
back: olive-brown back with streaks  
wings: olive-brown with white-edged feathers  
legs: long and skinny, with blackish coloring  
tail: short and dark, with white outer feathers

M&V's prediction: **prairie warbler** 0.002  
but we rejected it because...

0.696	crown: yellowish-green	0.000
0.688	forehead: yellow with black markings	0.309
0.722	nape: greenish-yellow	0.000
0.483	eyes: dark with thin white eye-ring	0.212
0.475	beak: small and pointed	0.149
0.672	throat: bright yellow	0.173
0.614	breast: bright yellow with faint streaks	0.551
0.624	belly: yellowish with light brown streaks	0.306
0.688	back: olive-green with faint streaks	0.100
0.575	wings: dark grayish-brown with white streaks	0.220
0.645	legs: pinkish-brown	0.000
0.699	tail: dark grayish-brown with white edges	0.142



Our prediction: **Beagle** 0.126  
because of the following...

head: round with a distinct dome shape, often a mix of white and brown or black fur  
ears: long, droopy and feathered, usually colored in rich brown or black, framing each side of the face  
muzzle: short and slightly tapered, covered in short brown, black, or white fur, with a black nose at the end  
body: compact and muscular, covered in a silky, wavy coat that can be a mix of white, brown, black and tan  
legs: short to medium length and straight, with feathered fur that matches the color of the body  
tail: medium length, often docked, covered in feathered fur, carried happily but never much above the level of the back

Top-2 prediction: **Papillon (Continental Toy Spaniel)** 0.023  
but we rejected it because...

0.671	head: round with a distinct "dome" shape, often covered in long, silky fur that can vary in color from black, brown, or white	0.474
0.497	ears: long, floppy, and heavily feathered, usually in deep chestnut brown or black, often hang down past the jawline	0.000
0.428	muzzle: short and tapered, usually the same color as the body fur with a black or brown nose at the end	0.014
0.200	body: compact and well-balanced, covered in silky fur that can be a blend of white, black, and brown	0.207
0.637	legs: short and straight, often covered in feathered fur that matches the body color, paws are small and compact	0.290
0.641	tail: medium-length, often covered in feathered fur, usually carried aloft but not above the level of the back	0.425

**Predicted**

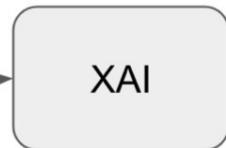
**Shape & size:** goose-sized or larger

**Color pattern:** brown, white fur

**Behavior:** Bird walking in shallow water and holding food

**Habitat:** wetland, shallow water

?

**Cornell BirdID description**

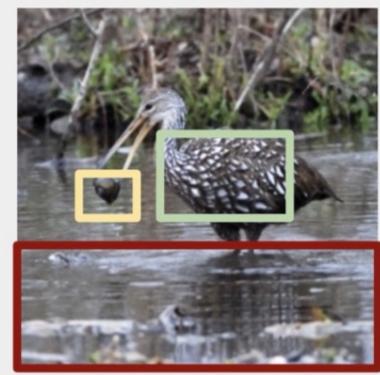
**Shape & size:** goose-sized or larger

**Color pattern:** Uniformly rich brown with heavy white spangling in the neck, back, and wing coverts

**Behavior:** Limpkins stalk snails by slowly walking through shallow water or on top of floating vegetation. Limpkins eat almost exclusively apple snails, plus at least three other native freshwater snail species and five species of freshwater mussels.

**Habitat:** Limpkins inhabit freshwater wetlands but can sometimes be found foraging in agricultural areas such as cane fields or in canals and ditches.

*Nearest neighbors from **Limpkin** class*



...



# Future XAI

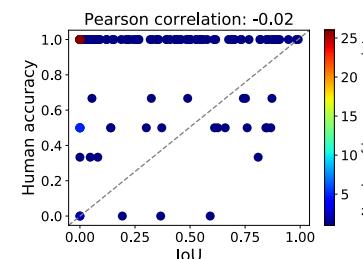
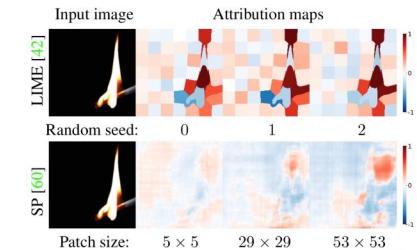
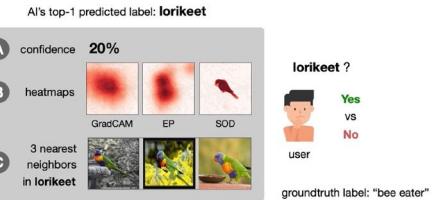
# Thank you!

# Conclusion

1. Build AIs that are accurate in edge cases
2. Build AIs that maximize human-AI team accuracy
3. Build AIs that humans can debug and edit

Trustworthy AI

Explainable AI



(b) Explar

Research gratefully funded by



cloud credit  
for research



CAREER  
AWARDS