# About my research

Giang Nguyen

# About me

🎓Giang Nguyen, pronounced Zi-ang, is a 3rd-year Ph.D. student at Auburn University, US.

🏃 He loves soccer ⚽, tennis 🎾, animals 🐈, and reading all kinds of things 📖.

🙇 He was/is fortunate to be advised by awesome people as shown!

**B.Eng** in
Electronics & Telecom,
HUST, Vietnam

**SWE** in
Dasan Networks,
Hanoi Vietnam

**M.Sc.** in
Computer Science,
KAIST, South Korea

**Ph.D. student** in
Computer Science,
Auburn University

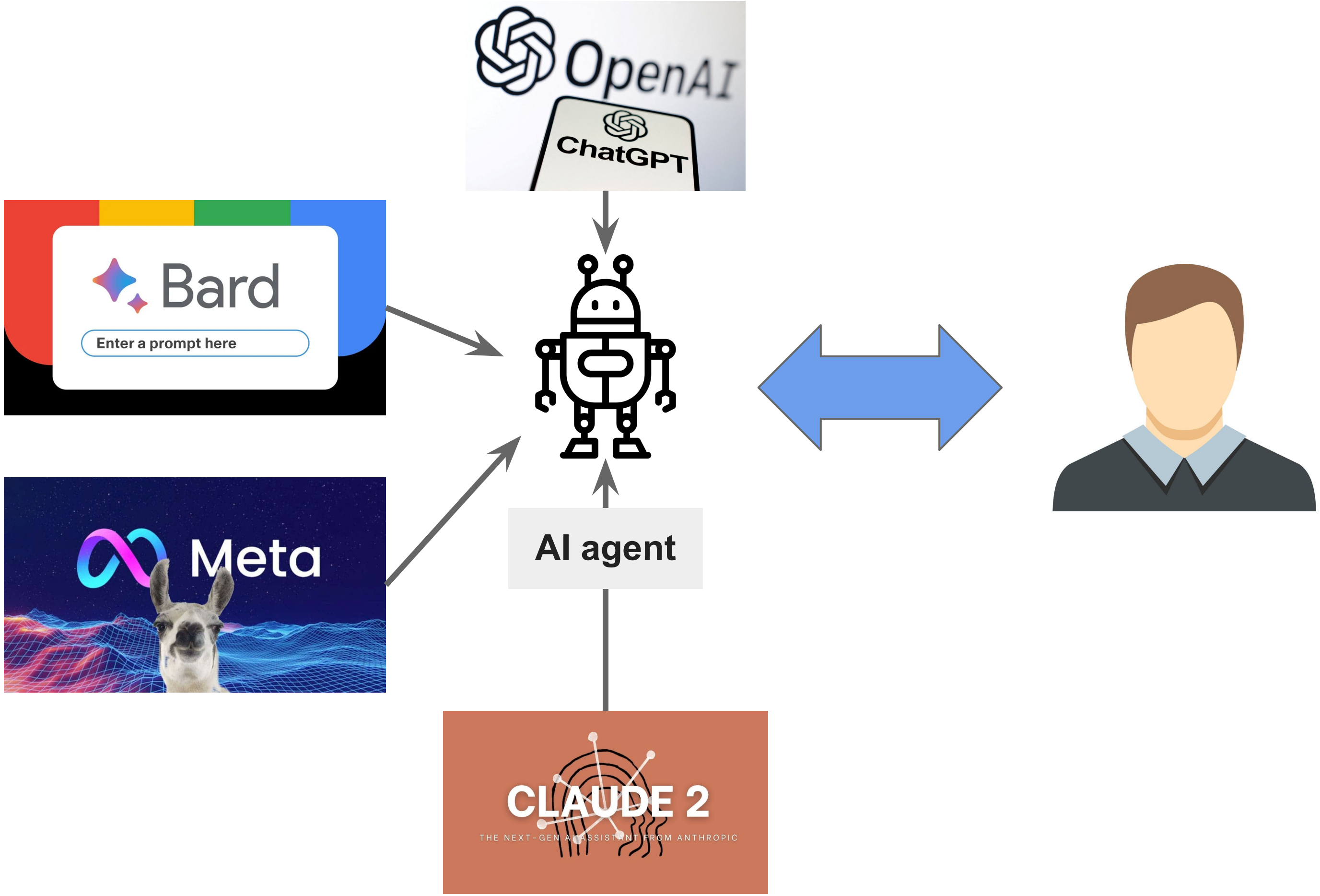
**2016**

**2018**

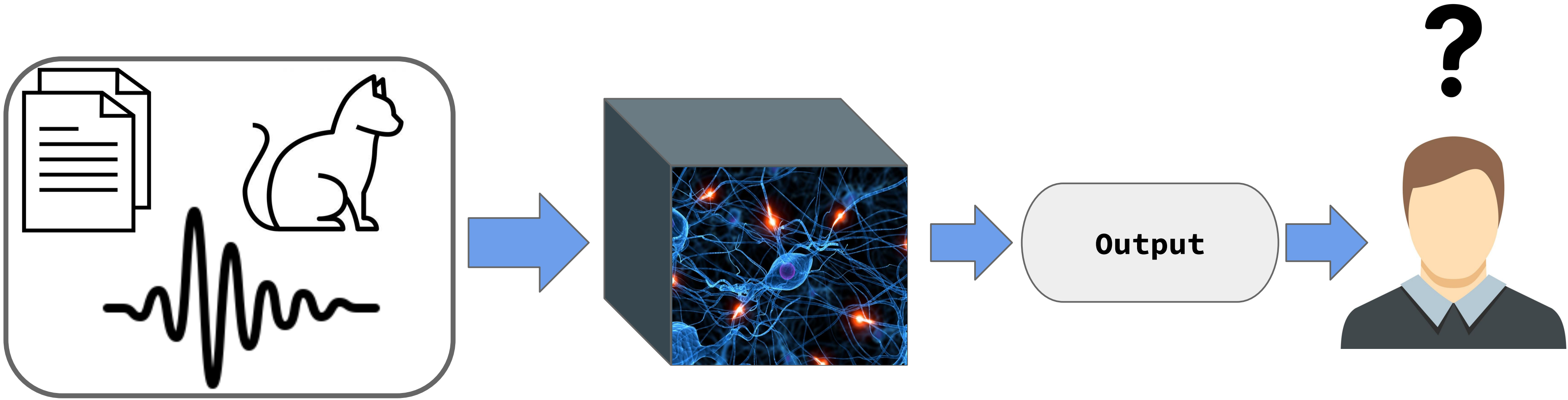
**2020**

**2021**

# Humans and AIs work together everyday



# Deep neural networks (AIs) are black boxes to humans



# Humans and AI working together effectively… via an **interface**

# Research #1:
# The effectiveness of feature attribution methods and its correlation with automatic evaluation scores, NeurIPS 2021.

Giang Nguyen, Daeyoung Kim, Anh Nguyen

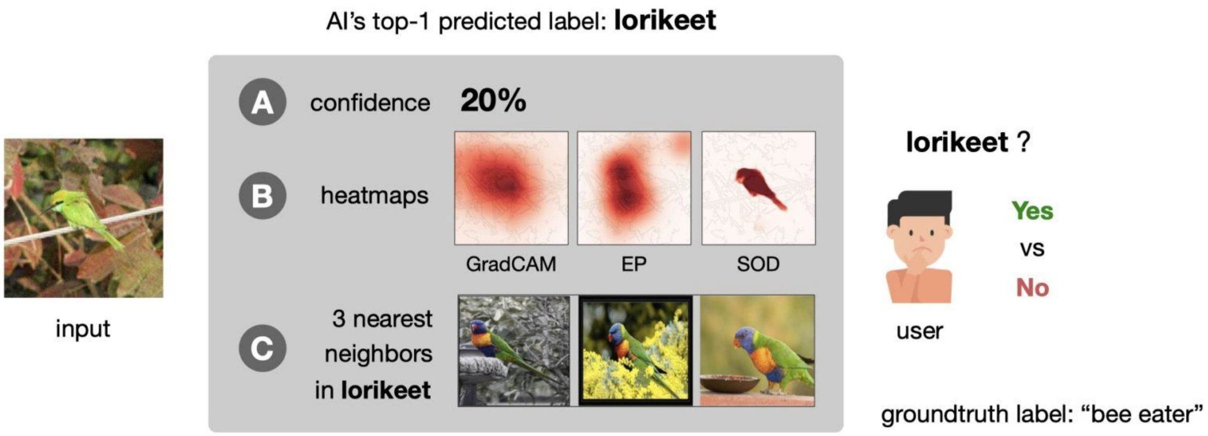
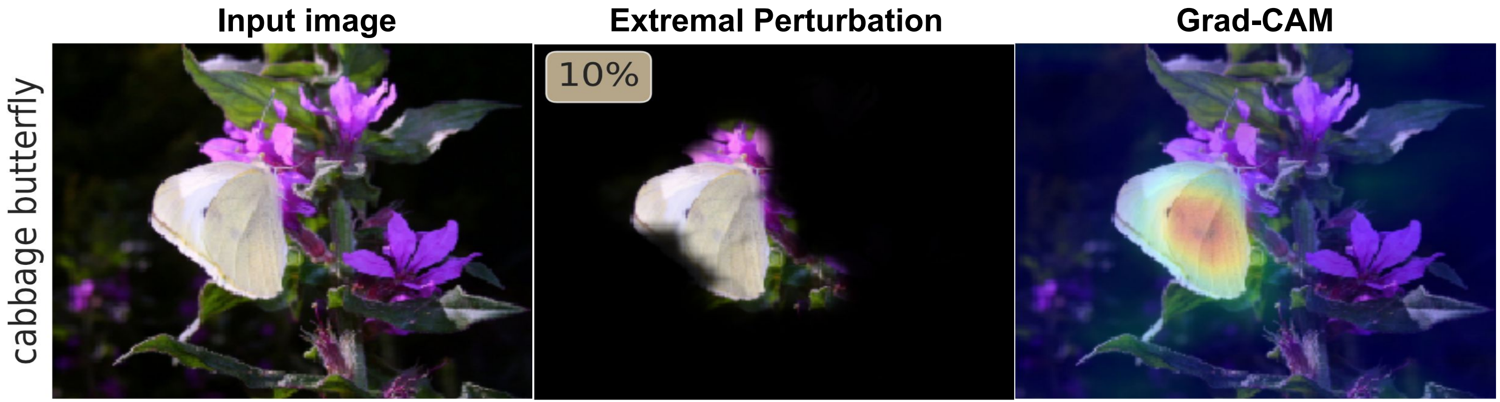



**Figure 1:** Given an input image, its top-1 predicted label (here, *lorikeet*) and confidence score (A), we asked the user to decide Yes or No whether the predicted label is accurate (here, the correct answer is No). The accuracy of users in this case is the performance of the human-AI team *without* visual explanations. We also compared this baseline with the treatments where *one* attribution map (B) or a set of three nearest neighbors (C) is also provided to the user (in addition to the confidence score).

# RQs

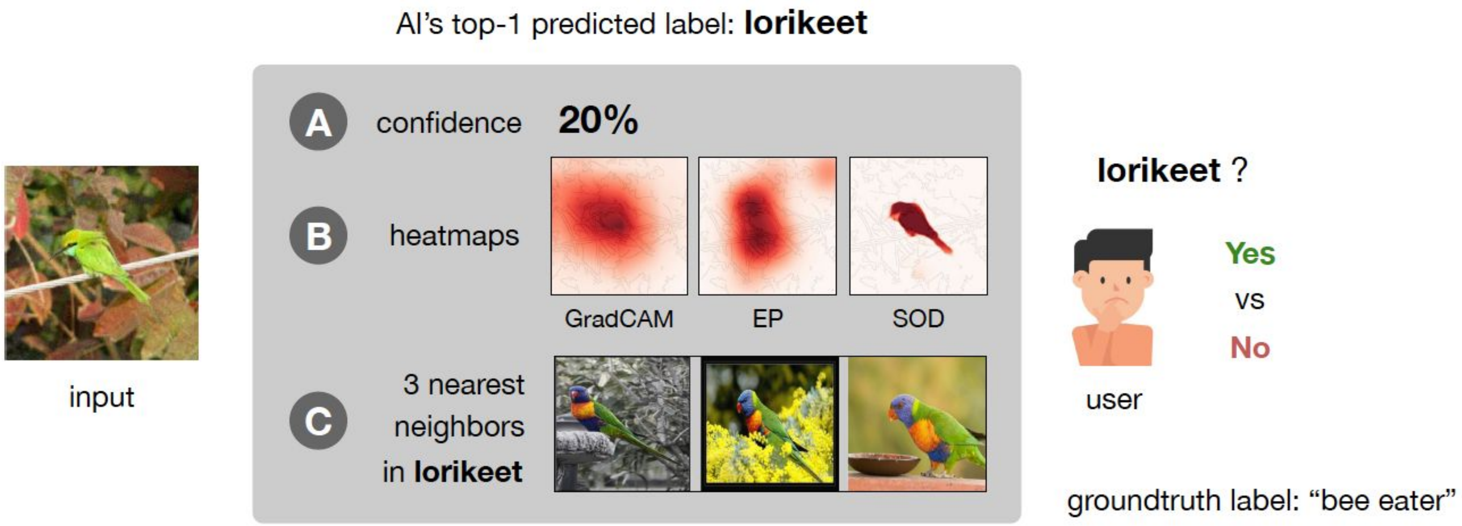



RQ1: Can existing popular XAI methods (AMs) help humans make better decisions when working with AI?

**Dozens** of attribution methods (AMs) have been tested on proxy benchmarks (insertion/deletion/IoU/pointing-game scores) rather than humans.

RQ2: Can an XAI method having high XAI scores help humans better?

# Experiment setup

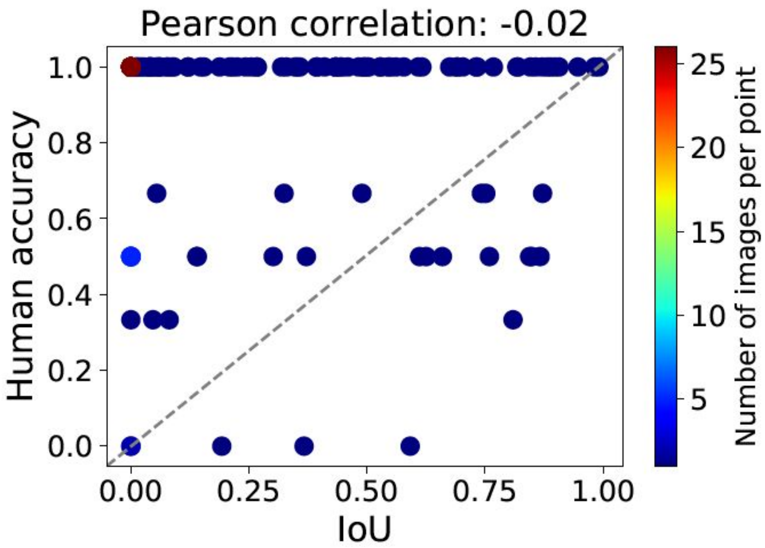



Setup: XAI methods help user inspect if AI is correct or wrong.

# Results

| Method | ImageNet | | Stanford Dogs | |
|---|---|---|---|---|
| | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| Confidence | 72.44 | 8.25 | **61.71** | 11.39 |
| GradCAM | 72.58 | 8.11 | 60.56 | 9.27 |
| EP | 73.85 | 6.88 | 56.67 | 10.57 |
| SOD | 72.06 | 7.63 | **61.67** | 10.87 |
| 3-NN | **76.08** | **5.86** | 57.20 | 10.58 |

1) AMs do not help users make better decisions. Rather, showing nearest-neighbor (NN) examples or not showing explanations aat all is better.

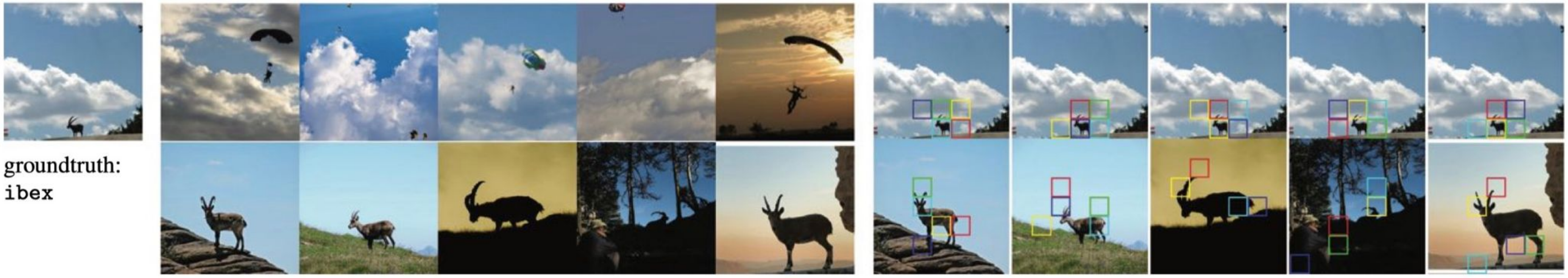



2) Evaluation metrics do not positively correlate with downstream utility in decision making.

# Research #2:
## Visual correspondence-based explanations improve AI robustness and human-AI team accuracy, NeurIPS 2022.

Giang Nguyen*, Mohammad Reza Taesiri*, Anh Nguyen

*co-first authors




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

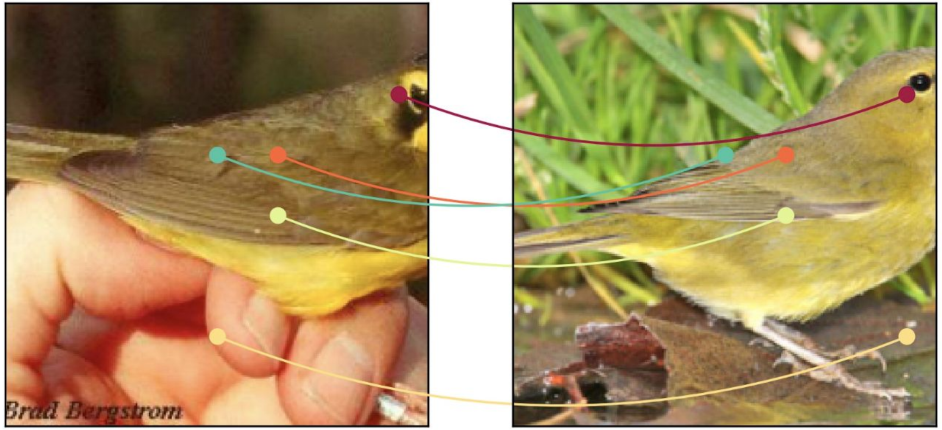
(b) Explanations for CHM-Corr's ibex decision

**Figure 1:** The ibex image is misclassified into parachute due to its similarity (clouds in blue sky) to parachute scenes (a). In contrast, CHM-Corr correctly labels the input as it matches ibex images mostly using the animal's features, discarding the background information (b).

# RQs

**Given that NN explanations are intuitive and help humans make better decisions.**

RQ1: How can we advance example-based explanations (NNs)?

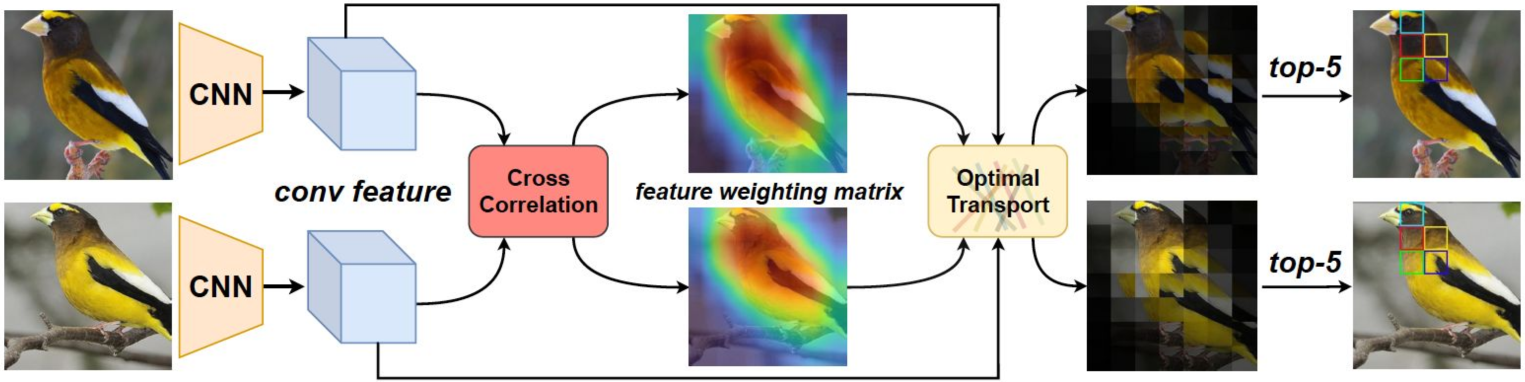



For humans, when comparing two objects, we leverage feature-to-feature comparisons or called correspondences. This explanation combine advantages of both AMs and NNs.

1. Showing extra information beyond input sample.
2. Pinpointing AI’s attention

RQ2: How to make this explanation useful for AI accuracy and human-AI team accuracy?

# EMD-Corr classifier




How to devise the optimal transport flow matrix?

1. Compute the similarities between two nodes in two images using cosine to get d_ij
2. Using CC to assign importance weight w_ij for each patch
3. Minimize the cost given the constraints of F and find the flow matrix F.
4. Find correspondences using coordinates of flow matrix

$$\text{Cost}(Q, G, \boldsymbol{F}) = \sum_{i=1}^{M} \sum_{j=1}^{M} d_{ij} f_{ij} \tag{2}$$

where $f_{ij} \geq 0$ and $\sum_{j=1}^{M} \sum_{i=1}^{M} f_{ij} = 1$. We use Eq. 1 to compute the ground distance $d_{ij}$ and run the Sinkhorn algorithm [21] for 100 iterations to seek the *optimal transport plan* $\boldsymbol{F}$. To assign importance weights (i.e., $w_{q_i}$ and $w_{g_j}$), we use cross-correlation (CC) maps from [68].

# Results

Table 1: Top-1 accuracy (%). ResNet-50 models' classification layer is fine-tuned on a specified training set in (b). All other classifiers are non-parametric, nearest-neighbor models based on pretrained ResNet-50 features (a) and retrieve neighbors from the training set (b) during testing. EMD-Corr & CHM-Corr outperform ResNet-50 models on all OOD datasets (e.g. +4.39 on Adversarial Patch) and slightly underperform on in-distribution sets (e.g. -0.72 on ImageNet-ReaL).

| Test set | Features (a) | Training set (b) | ResNet-50 | kNN | EMD-Corr | CHM-Corr | CHM-Corr+ |
|---|---|---|---|---|---|---|---|
| ImageNet [63] | ImageNet | ImageNet | **76.13** | 74.77 | 74.93 (-1.20) | 74.40 (-1.73) | n/a |
| ImageNet-ReaL [14] | ImageNet | ImageNet | **83.04** | 82.05 | 82.32 (-0.72) | 81.97 (-1.07) | n/a |
| ImageNet-R [35] | ImageNet | ImageNet | 36.17 | 36.18 | **37.75** (+1.58) | 37.62 (+1.45) | n/a |
| ImageNet Sketch [72] | ImageNet | ImageNet | 24.09 | 24.72 | 25.36 (+1.27) | **25.61** (+1.52) | n/a |
| DAmageNet [18] | ImageNet | ImageNet | 5.93 | 7.59 | **8.16** (+2.23) | 8.10 (+2.17) | n/a |
| Adversarial Patch [15] | ImageNet | ImageNet | 55.04 | 59.30 | 59.43 (+4.39) | **59.86** (+4.82) | n/a |
| CUB [71] | ImageNet | CUB | n/a | 54.72 | **60.29** | 53.65 | 49.63 |
| CUB [71] | iNaturalist [70] | CUB | **85.83** | 85.46 | 84.98 (-0.85) | 83.27 (-2.56) | 81.54 |

1) EMD-Corr improves AI robustness

Table 2: Human-only accuracy (%)

| Method | ImageNet-ReaL | | CUB | |
|---|---|---|---|---|
| | Users | Accuracy | Users | Accuracy |
| ResNet-50 | 60 | **81.56** ± 5.54 | 60 | 65.50 ± 7.46 |
| kNN | 59 | 75.76 ± 8.55 | 59 | 64.75 ± 7.14 |
| EMD-Corr | 59 | **78.87** ± 6.57 | 58 | **67.64** ± 7.44 |
| CHM-Corr | 59 | 77.23 ± 7.56 | 59 | **69.72** ± 9.08 |
| EMD-NN | 57 | 77.72 ± 8.27 | 59 | 64.12 ± 7.07 |
| CHM-NN | 60 | 77.56 ± 6.91 | 60 | 65.72 ± 8.14 |

Table 3: AI-only and Human-AI team accuracy (%)

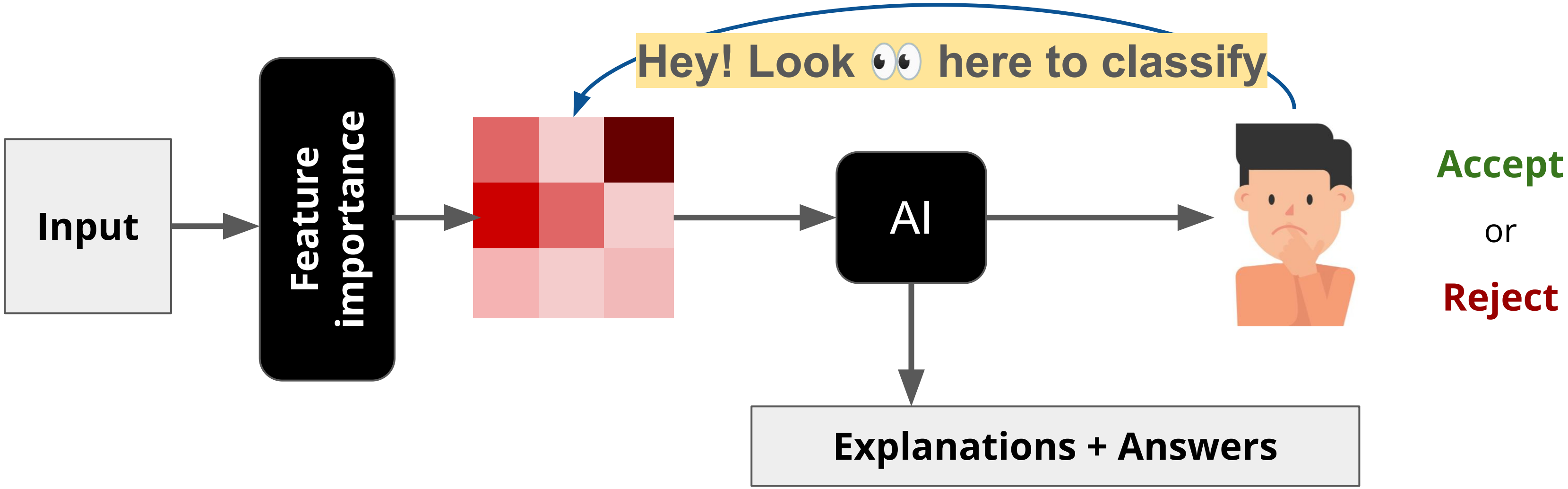
| Method | ImageNet-ReaL | | CUB | |
|---|---|---|---|---|
| | AI-only | Human-AI | AI-only | Human-AI |
| ResNet-50 | 86.05 | 90.41 (+4.36) | 87.11 | 87.74 (+0.63) |
| kNN | 85.95 | 87.85 (+1.90) | 87.40 | 86.56 (-0.84) |
| EMD-Corr | 85.91 | 89.48 (+3.57) | 86.88 | 87.03 (+0.15) |
| CHM-Corr | 85.36 | 88.51 (+3.15) | 85.48 | 87.22 (+1.74) |
| *mean* | 85.18 | 89.06 (+3.88) | 86.18 | 87.14 (+0.96) |

2) Our explanations improve both human and human-AI team accuracy.

# Research #3:

## “Allowing humans to interactively guide machines where to look”

*does not always* improve human-AI team’s classification accuracy



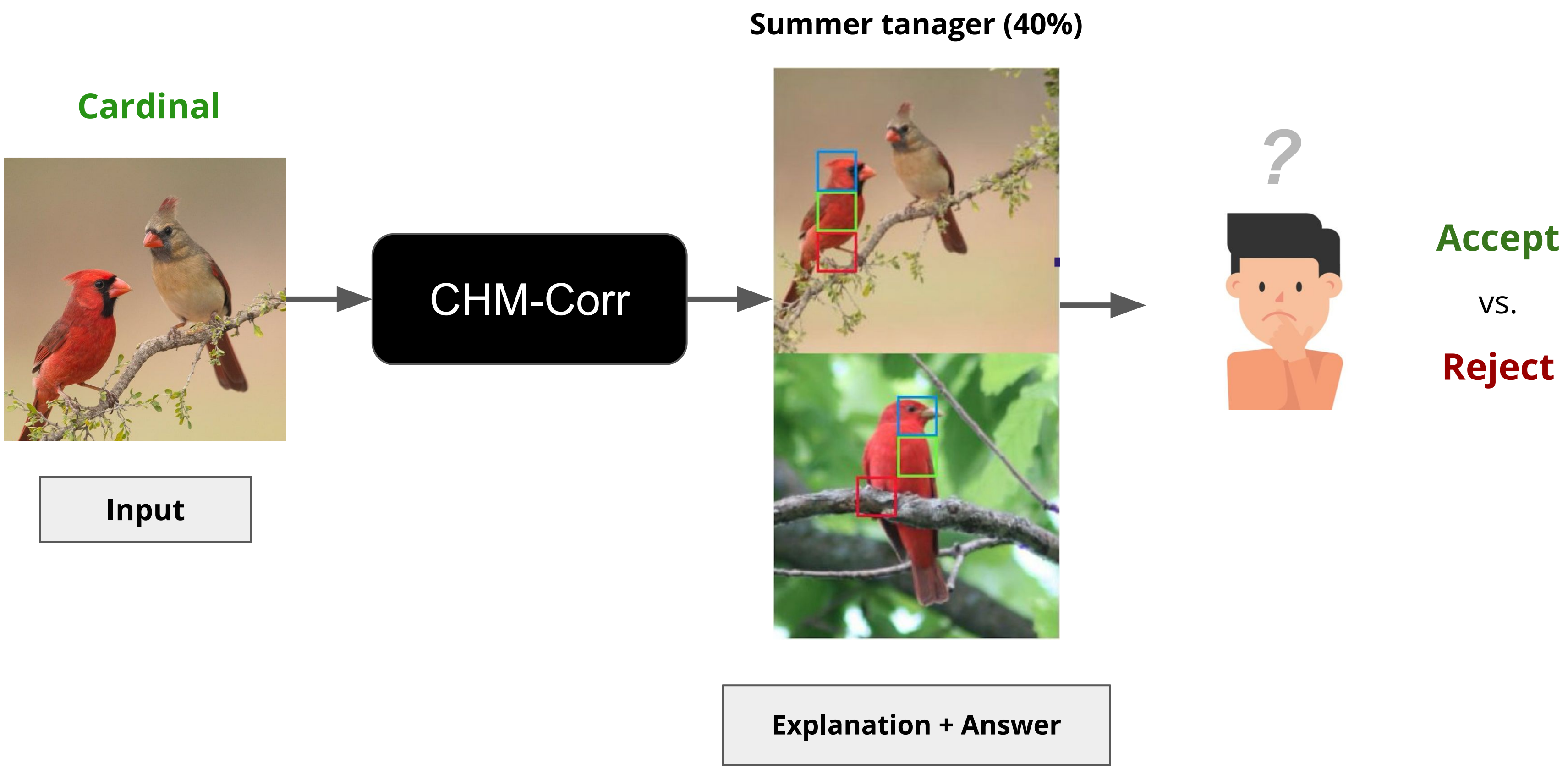


Giang Nguyen, Mohammad Reza Taesiri, Sunnie S. Y. Kim, Anh Nguyen

AUBURN UNIVERSITY
UNIVERSITY OF ALBERTA
PRINCETON UNIVERSITY


CVPR
JUNE 17-21, 2024
SEATTLE, WA

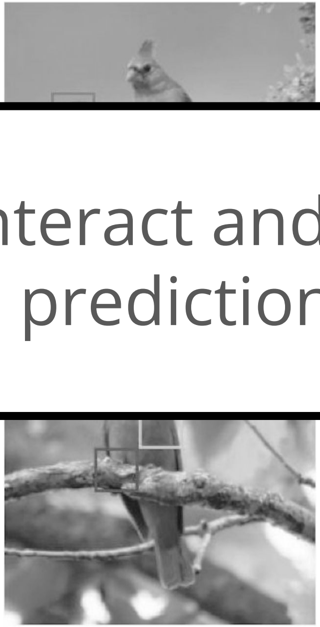
# State-of-the-art explanations are static and limit human understanding



# State-of-the-art explanations are static and limit human understanding

Summer tanager (40%)

Cardinal



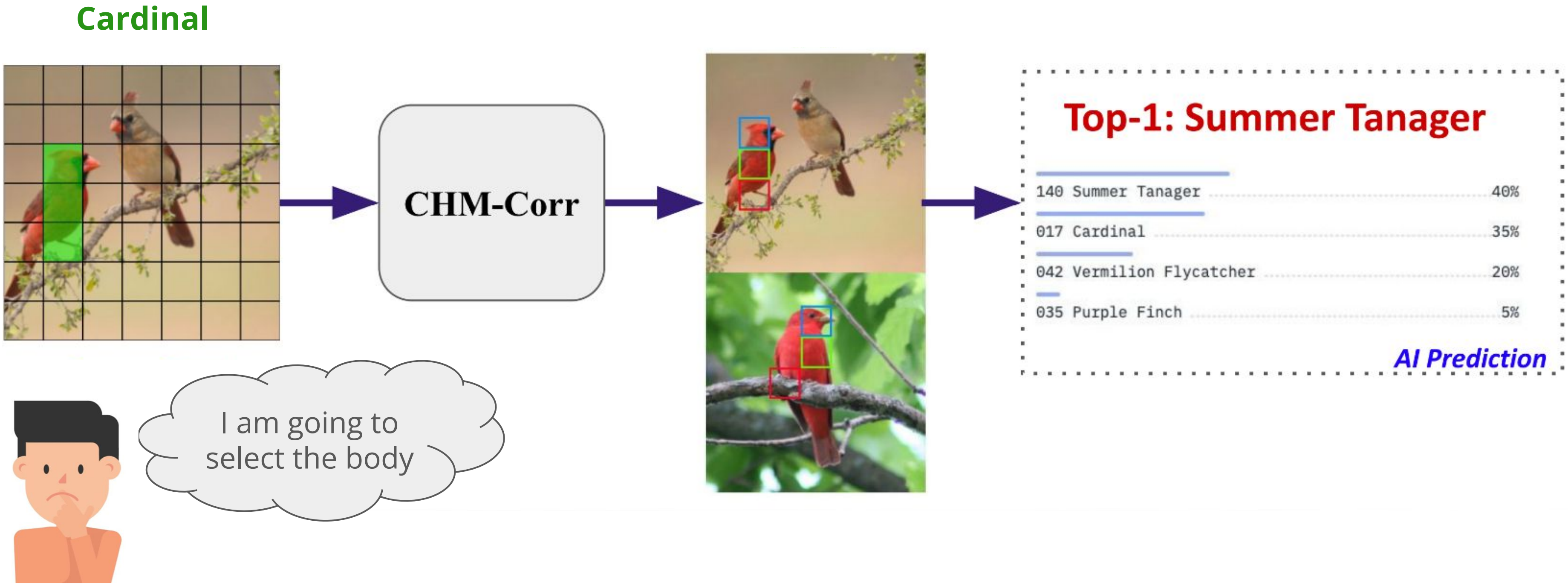
What if we allow users to interact and manipulate the AI's attention to generate more predictions and explanations?

Input

Explanation + Answer

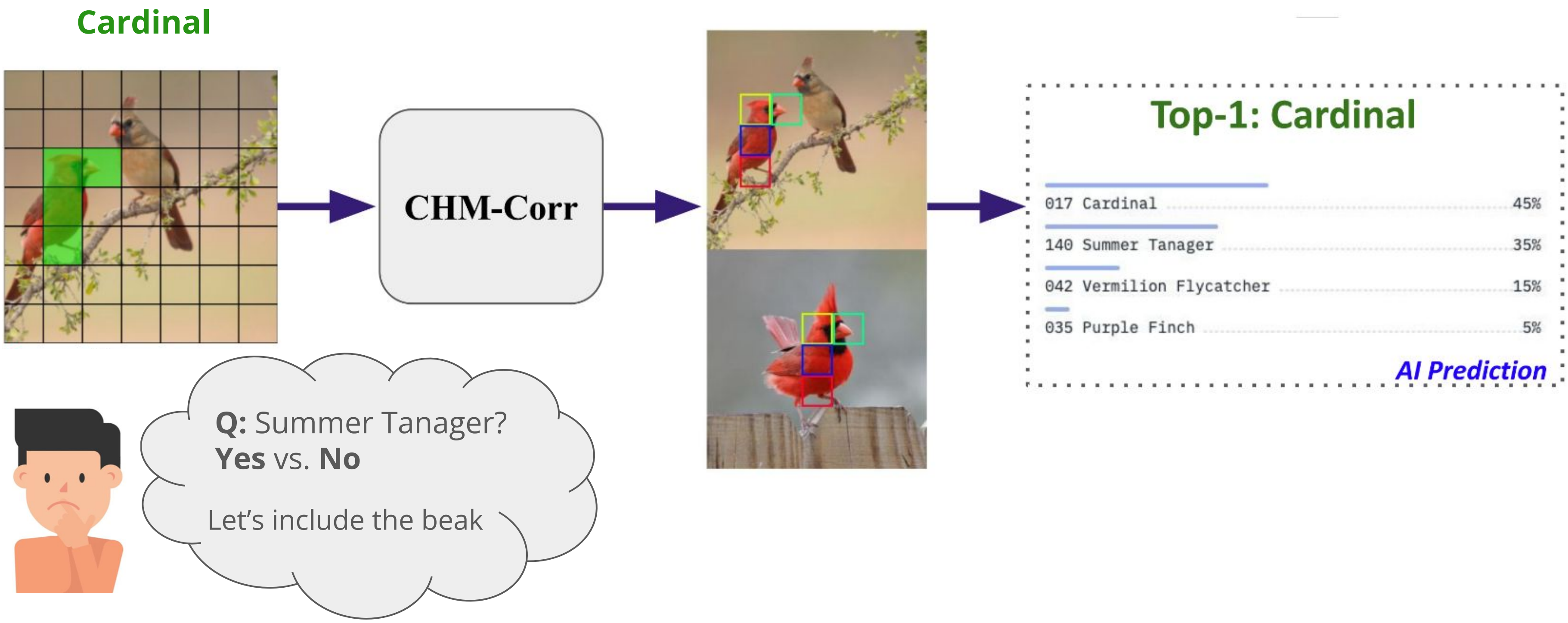
# Interactively editing model attention help users gain insights into: *if*, *when*, and *how* the model changes its predictions

## Step 1

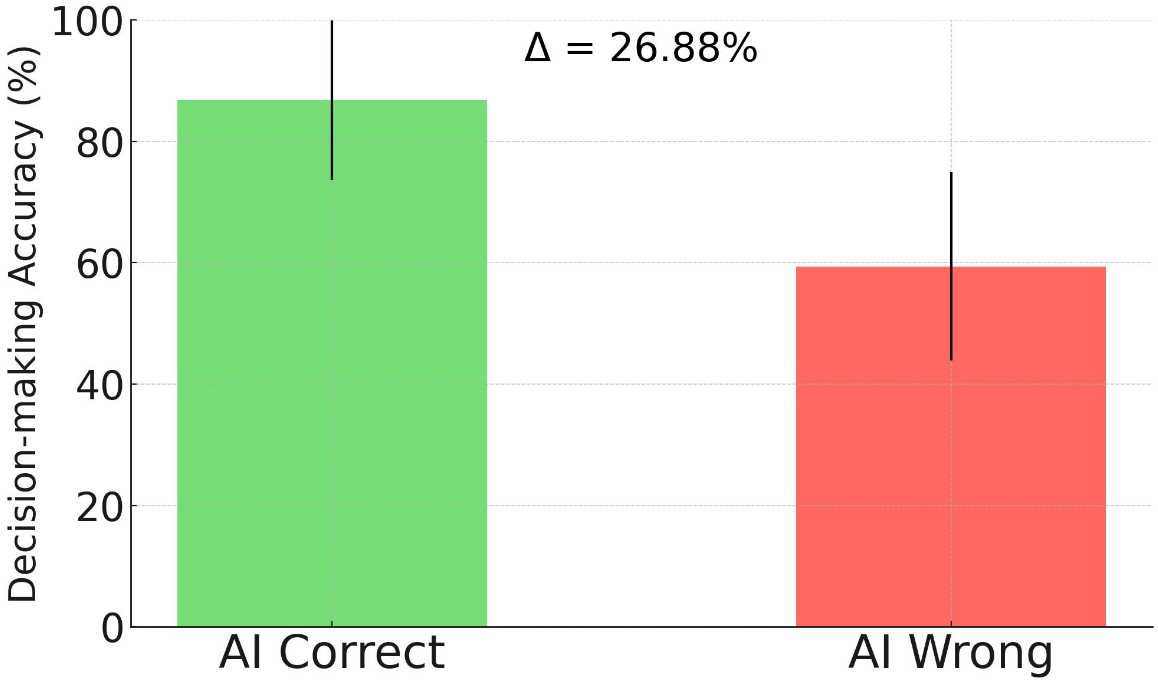


# Interactively editing model attention help users gain insights into: *if*, *when*, and *how* the model changes its predictions

## Step 2

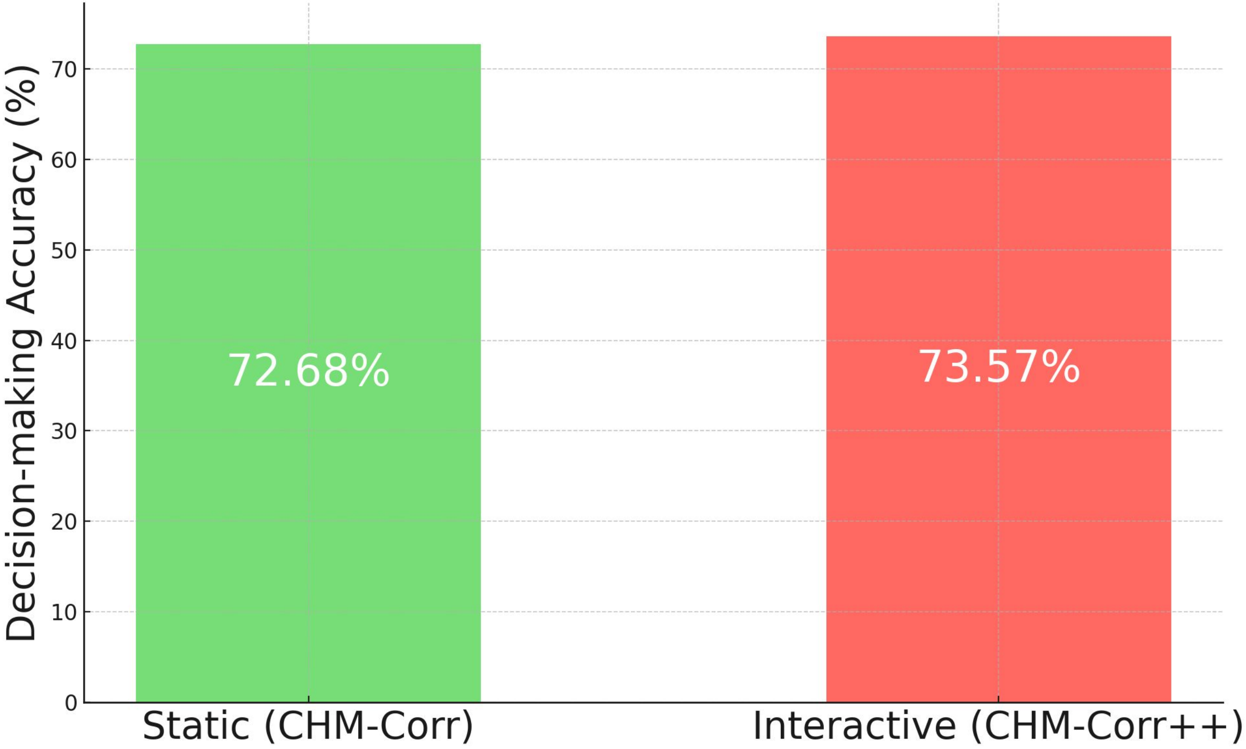


# Despite interactivity, it is still challenging to detect when AI is wrong



# We thought it would, but unfortunately NO!

***Interactivity does not improve human decision-making accuracy***



# Final Remarks

**Paper**: arxiv.org/pdf/2404.05238

**Demo**: 137.184.82.109:7080

**Code**: github.com/anguyen8/chm-corr-interactive

**Give it a try @**


Giang

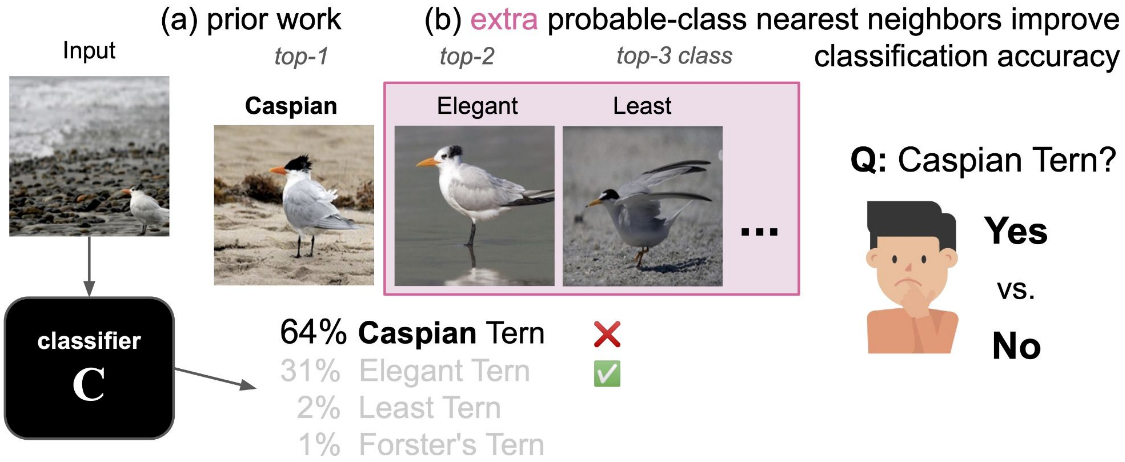

Mohammad Reza


Sunnie

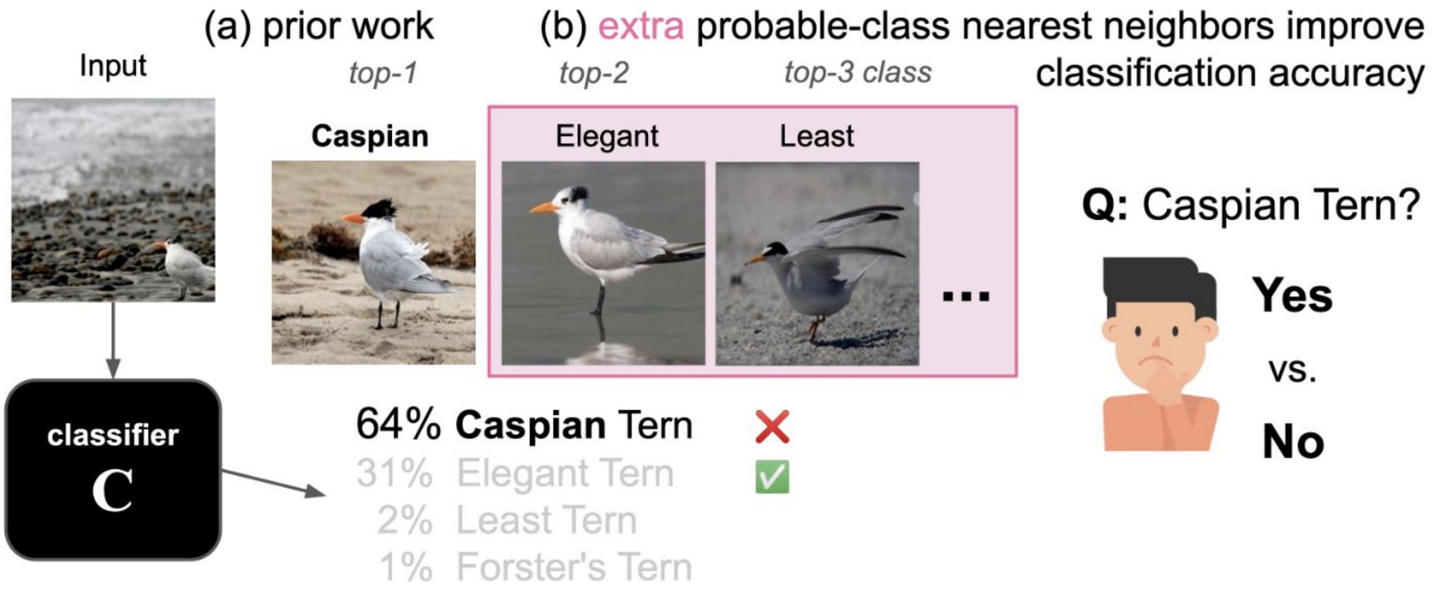

Anh

# Research #4:
# PCNN: Probable-Class Nearest-Neighbor Explanations Improve Fine-Grained Image Classification Accuracy for AIs and Humans, TMLR2024.

Giang Nguyen, Valerie Chen, Mohammad Taesiri, Anh Nguyen

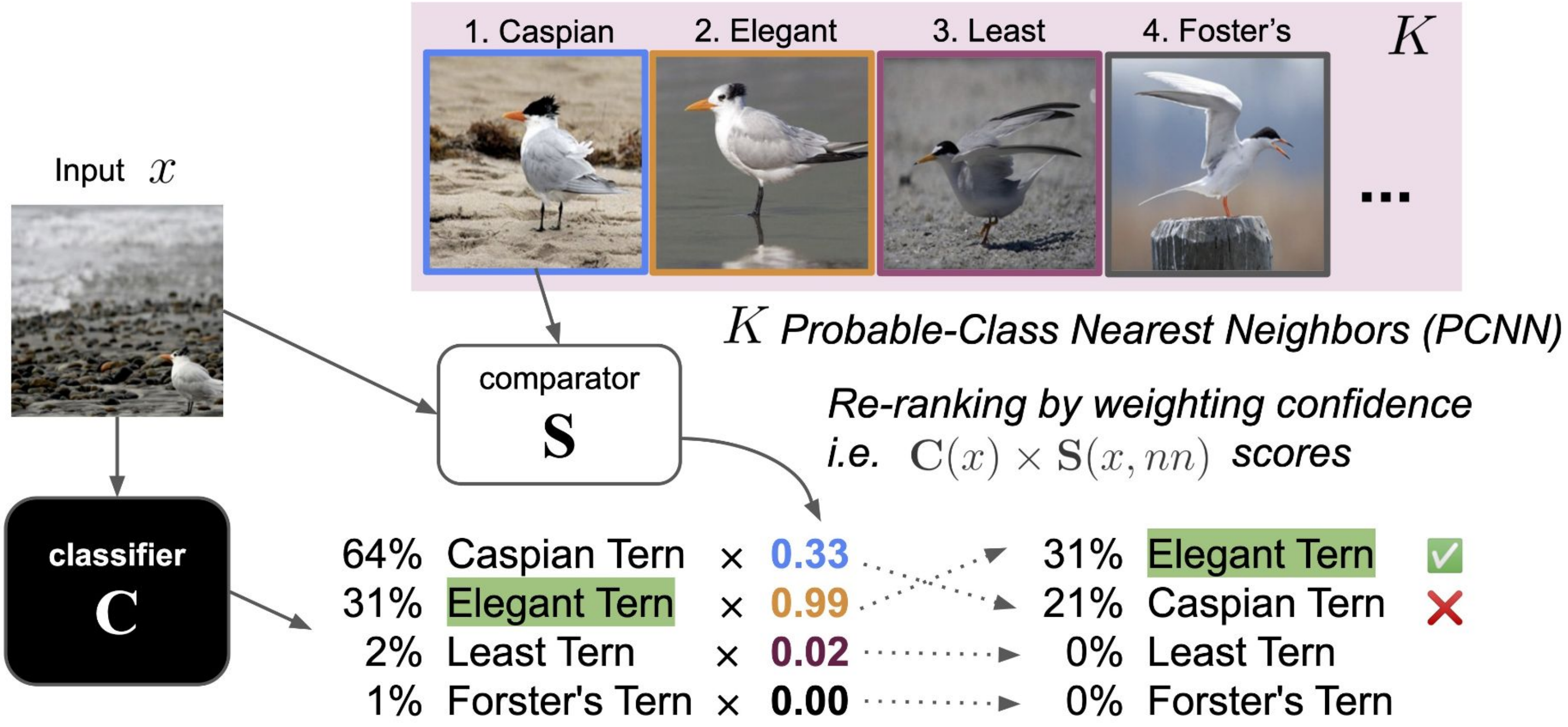


# Motivation

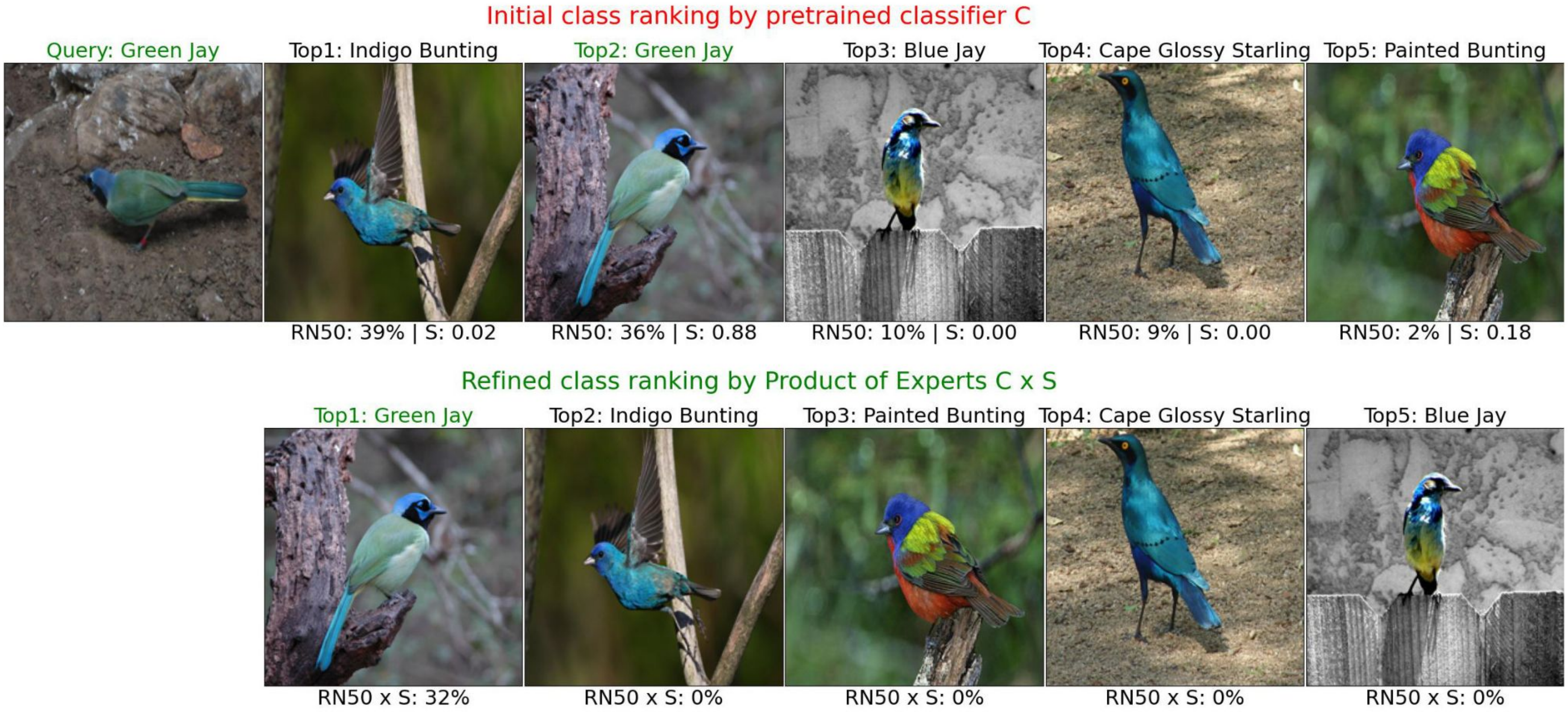



Given an input image $x$ and a black-box, pretrained classifier $C$ that predicts the label for $x$. Prior works (a) often show only the nearest neighbors from the top-1 predicted class as explanations for the decision, which often *fools* humans into accepting *wrong* decisions (here, **Caspian Tern**) due to the similarity between the input and top-1 class examples. Instead, including extra nearest neighbors (b) from top-2 to top-$K$ classes improves not only human accuracy on this binary distinction task but also AI's accuracy on standard fine-grained image classification tasks (see how below).

# A novel reranking-based algorithm

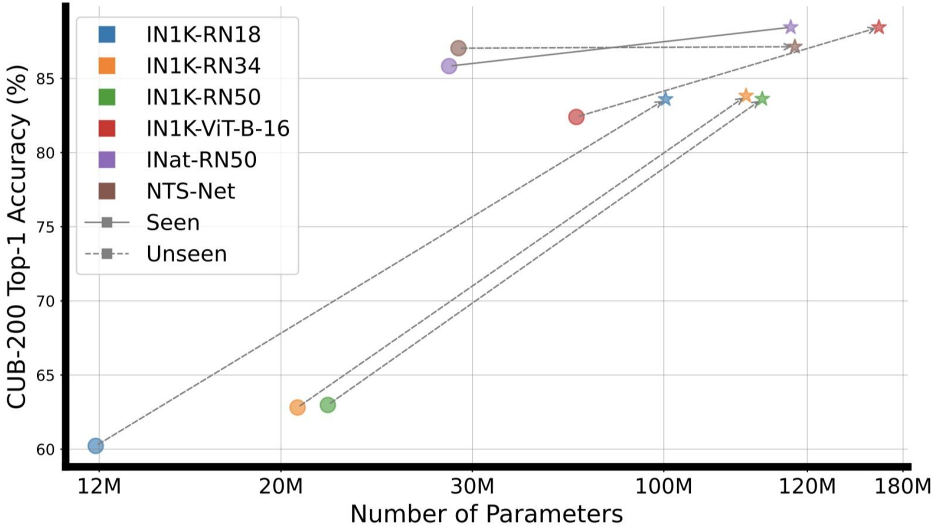


# Reranking samples

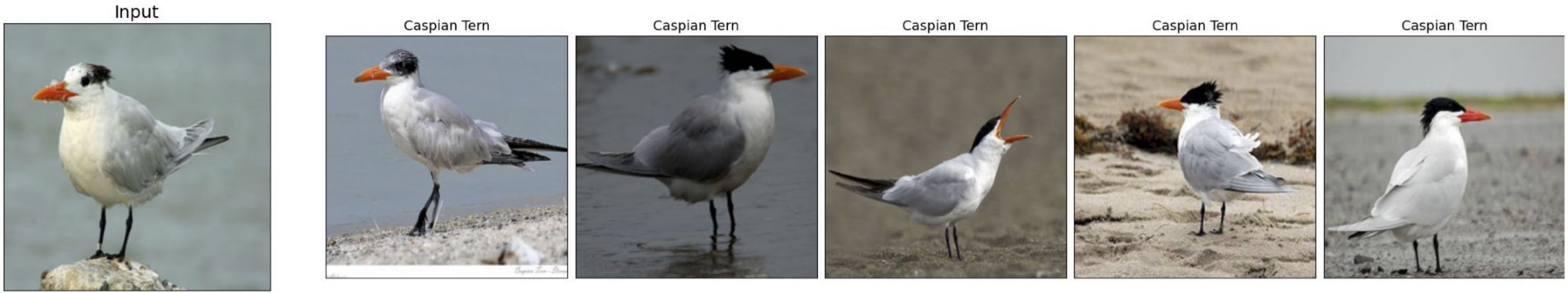


# Results – Explanations help improve AI accuracy

| Dataset | Pre-trained | RN18 | RN18 × S | RN34 | RN34 × S | RN50 | RN50 × S |
|---|---|---|---|---|---|---|---|
| CUB-200 | iNaturalist | N/A | N/A | N/A | N/A | 85.83 | 88.59 ( +2.76 ) |
| | ImageNet | 60.22 | 71.09 ( +10.87 ) | 62.81 | 74.59 ( +11.78 ) | 62.98 | 74.46 ( +11.48 ) |
| Cars-196 | ImageNet | 86.17 | 88.27 ( +2.10 ) | 82.99 | 86.02 ( +3.03 ) | 89.73 | 91.06 ( +1.33 ) |
| Dogs-120 | ImageNet | 78.75 | 79.58 ( +0.83 ) | 82.58 | 83.62 ( +1.04 ) | 85.82 | 86.31 ( +0.49 ) |

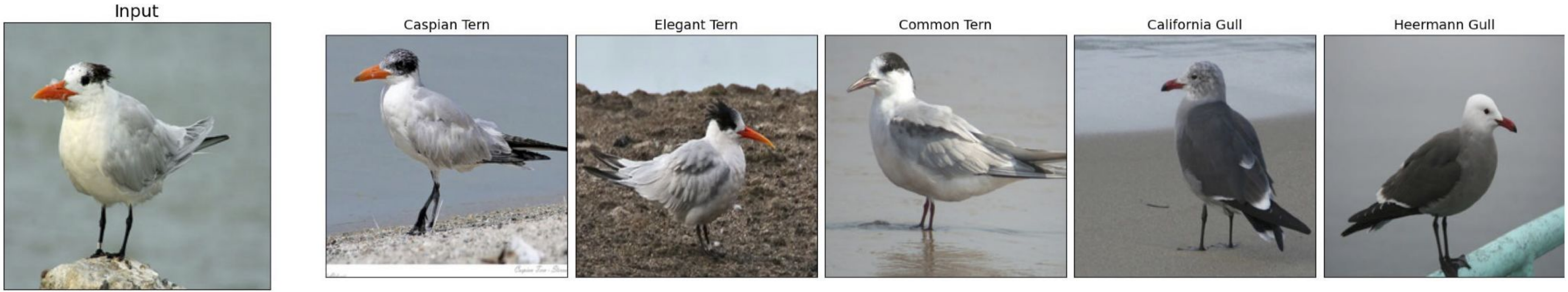


# Results – Explanations help Humans understand AIs




YES

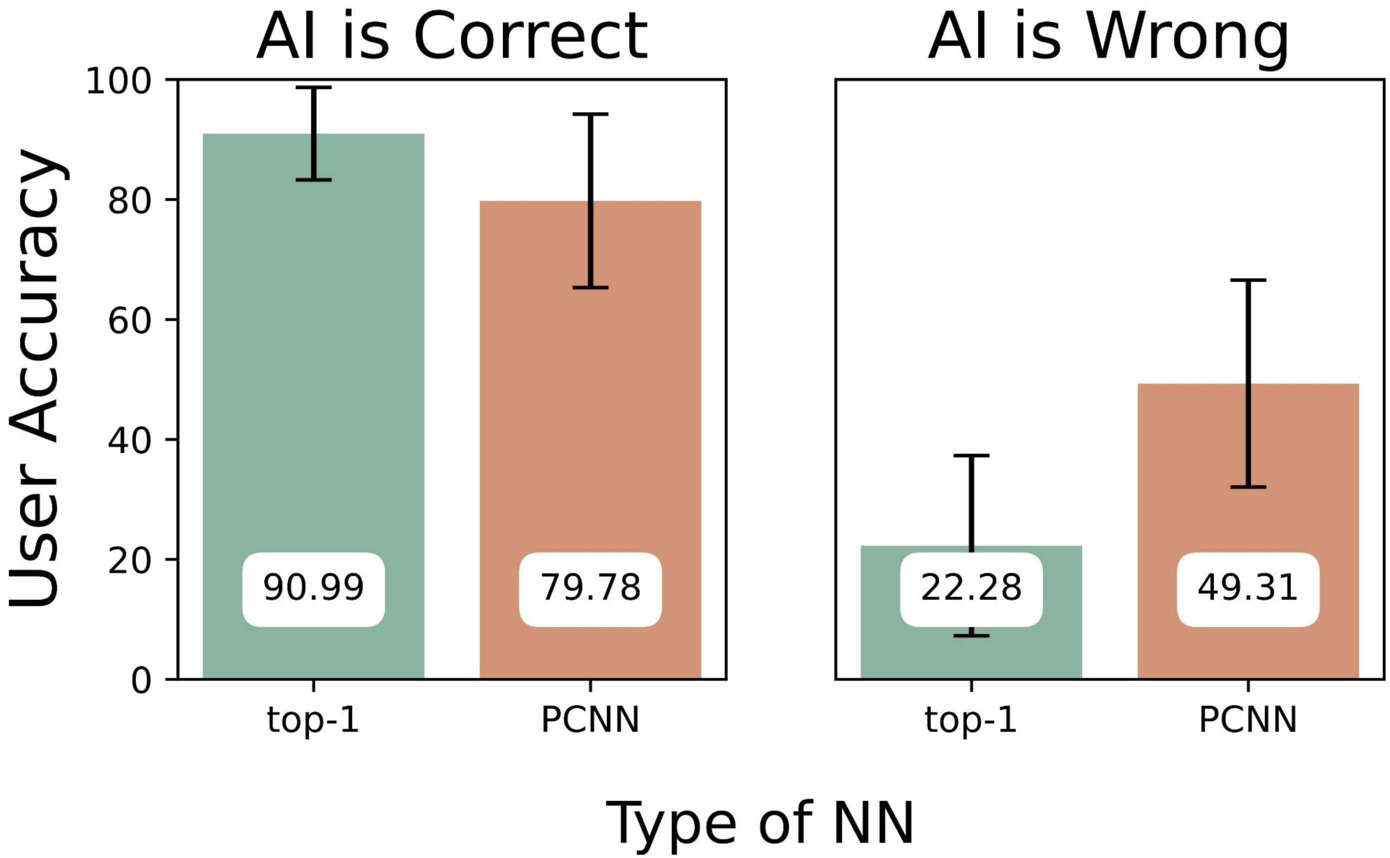
NO




YES

NO

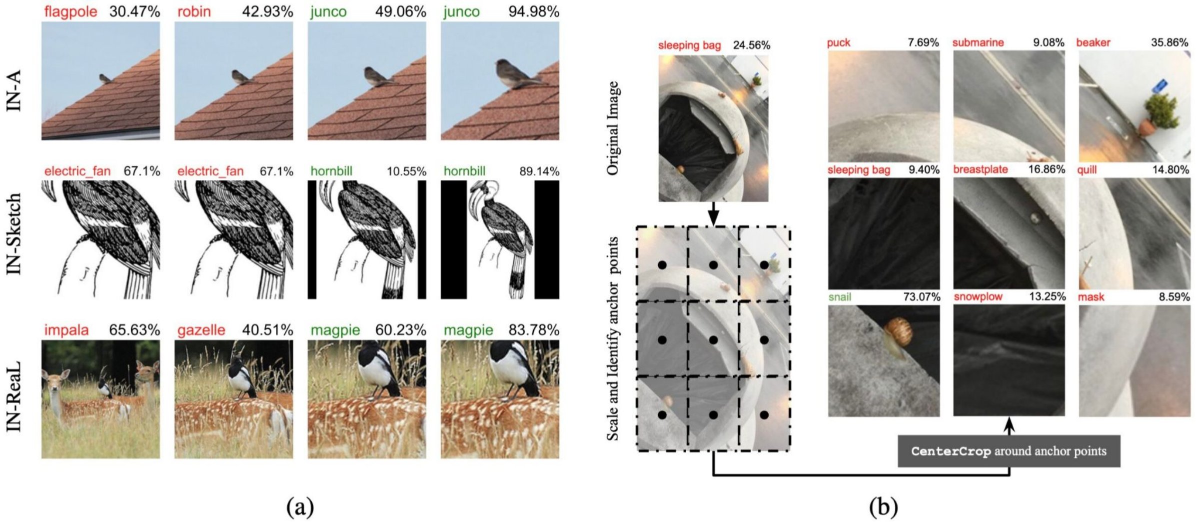
# Results – Explanations help Humans understand AIs



# Research #5:

## ImageNet-Hard: The Hardest Images Remaining from a Study of the Power of Zoom and Spatial Biases in Image Classification, NeurIPS 2023.

Mohammad Reza Taesiri, Giang Nguyen, Sarra Habchi, Cor-Paul Bezemer, Anh Nguyen



# RQs

**Current best image classifiers can score > 90% on ImageNet.**

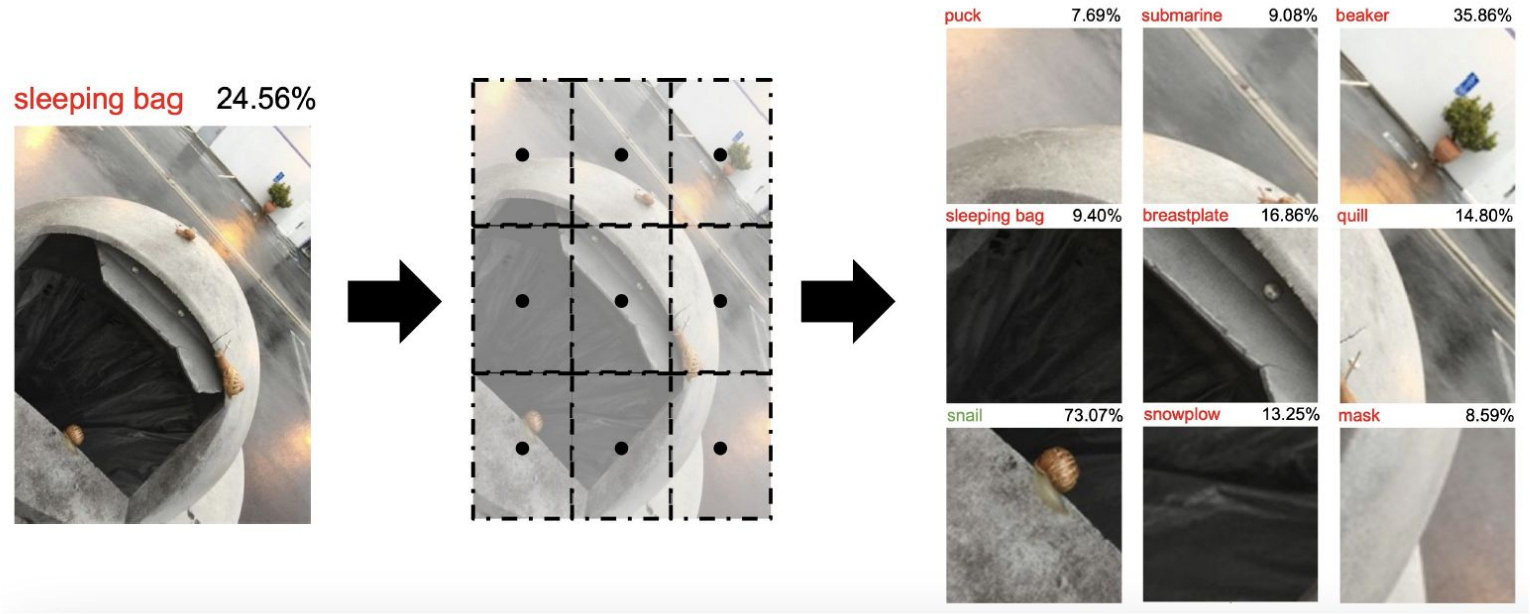
RQ1: What makes image classifiers so good since AlexNet (2012)?

RQ2: Are image classification benchmarks biased towards the center (the common practice in image classification)?

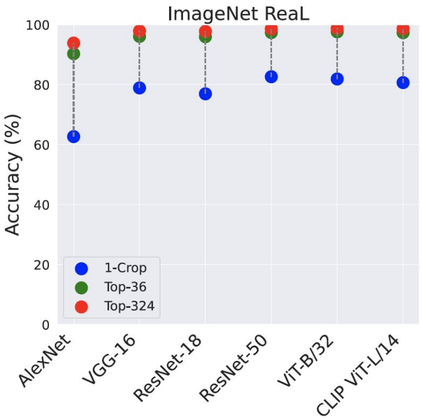
RQ3: If Zooming is the driving force (winning factor), can we have a dataset that challenges Zooming?

# Method

We approach the problem from the Zooming perspectives.



# Results




| | | |
|---|---|---|
| 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) |

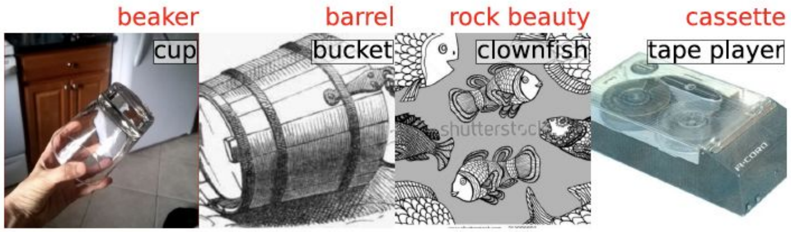
ImageNet-ReaL

| | | |
|---|---|---|
| 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) |

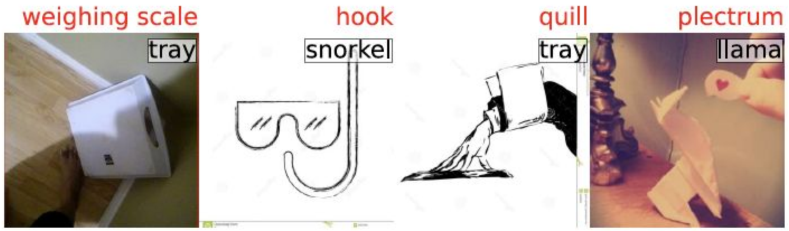
ImageNet-A

1) Representation learning is good enough since 2012 😱

2) ImageNet-A and ObjectNet are highly biased.




***Common*** misclassifications *(40%)*




***Rare*** misclassifications *(60%)*

3) Introducing ImageNet-Hard: A dataset with ~11K images that remain unclassifiable after many classification attempts at various zoom locations and crops.

# Summary of my research

1. **Building XAI methods (AI Interpretability)**

I am the author of explanation methods for computer vision systems: visual correspondences [2] (visual-corr) and probable-class nearest neighbors [5] (PCNN)

2. **Building Human-AI interaction (human in the loop via AI explanations)**

In 4 of my first-author papers written at Auburn, I tested how humans can work with AI via explanations to improve human decision-making performance [1,2,4,5]

3. **Making AI models robust (AI robustness)**

I introduced interpretable-by-design network [2] and a novel data augmentation techniques to make AI more robust against OOD samples [3]

## Selected Publications:


[1] The effectiveness of feature attribution methods and its correlation with automatic evaluation scores, NeurIPS'21.

[2] Visual correspondence-based explanations improve AI robustness and human-AI team accuracy, NeurIPS'22.

[3] ImageNet-Hard: The hardest images remaining from a study of the power of zoom and spatial biases in image classification, NeurIPS'23.

[4] Allowing humans to interactively guide machines where to look does not always improve a human-AI team's classification accuracy, CVPRW'24.

[5] PCNN: Probable-Class Nearest-Neighbor Explanations Improve Fine-Grained Image Classification Accuracy for AIs and Humans, TMLR'2024.