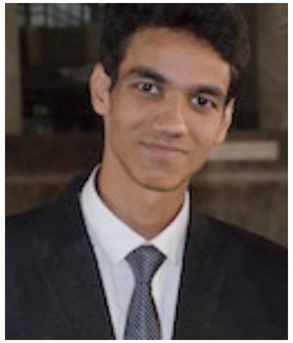
Counterfactual Visual Explanations



Yash Goyal
(Georgia Tech)

Ziyan Wu
(Siemens)

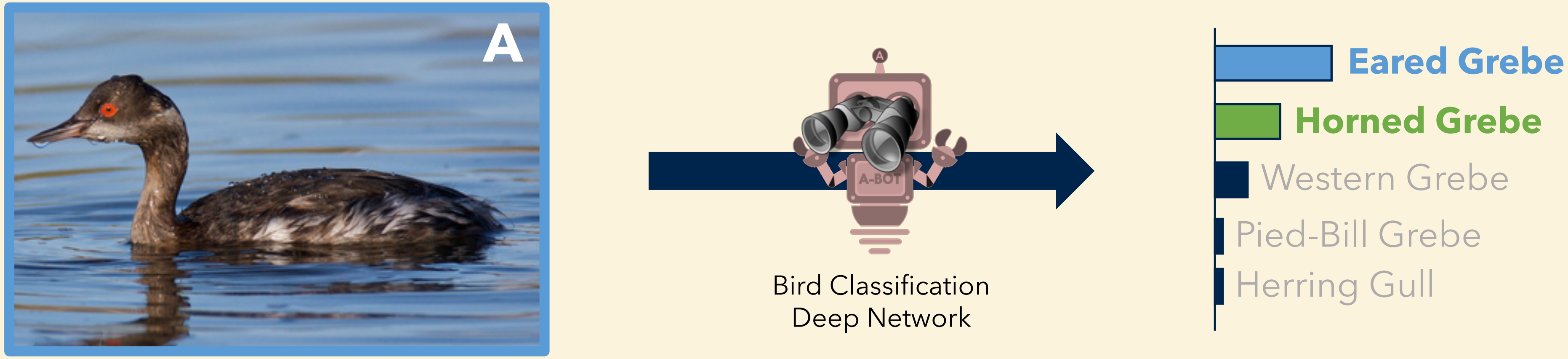
Jan Ernst
(Siemens)

Dhruv Batra
(Georgia Tech)

Devi Parikh
(Georgia Tech)

Stefan Lee
(Georgia Tech)

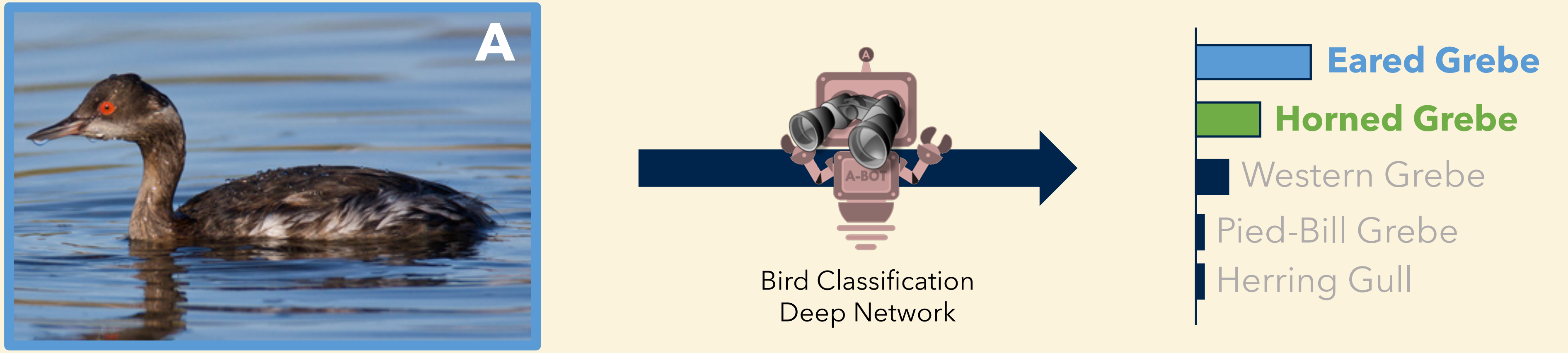
Counterfactual Visual Explanations



Bird Classification
Deep Network

An XAI Question: Why did the model predict **Eared Grebe** instead of **Horned Grebe**?

Counterfactual Visual Explanations

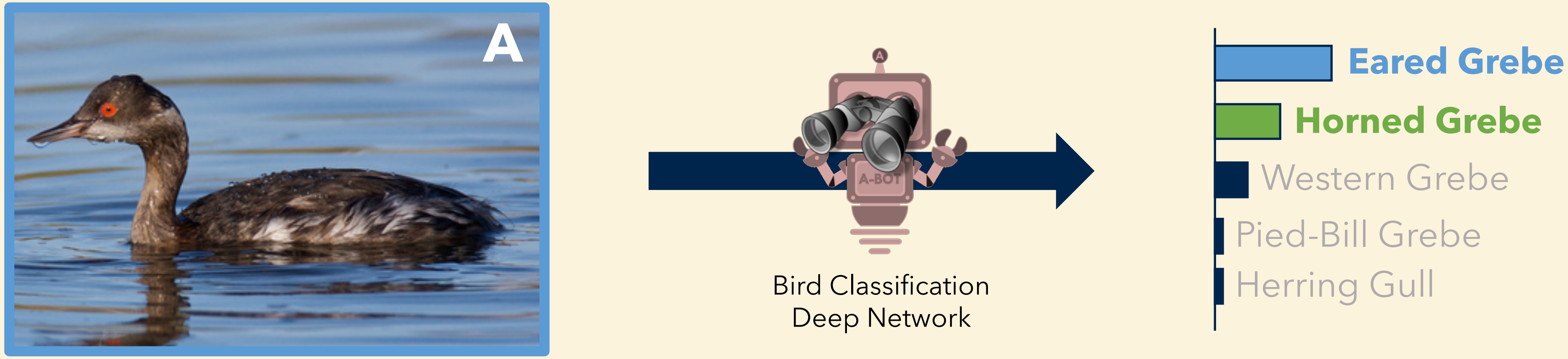


Bird Classification
Deep Network

An XAI Question: Why did the model predict **Eared Grebe** instead of **Horned Grebe**?

For input X, why did the model predict Y instead of Z?

Counterfactual Visual Explanations



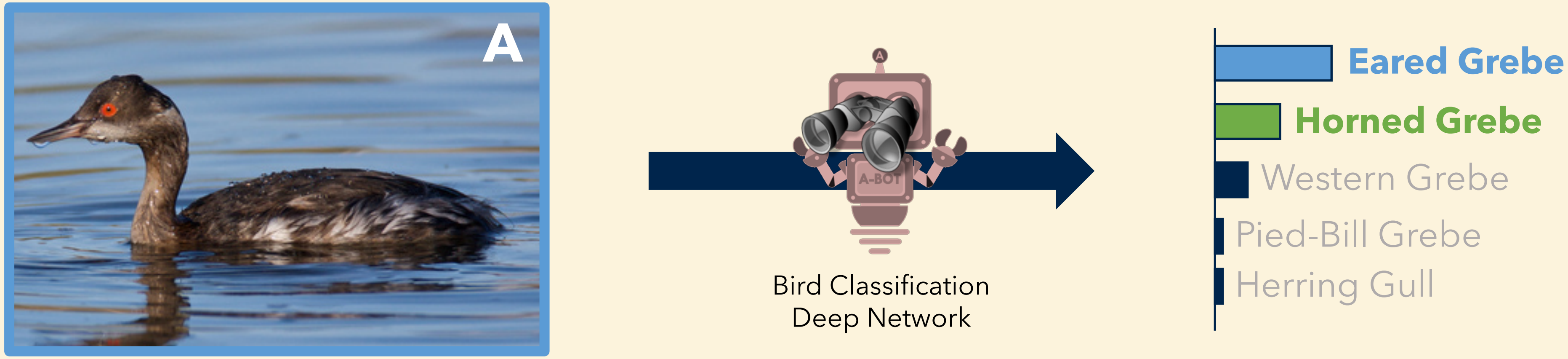
An XAI Question: Why did the model predict **Eared Grebe** instead of **Horned Grebe**?

For input X, why did the model predict Y instead of Z?

Explanation through Counterfactual:

If **X** was **X***, then the outcome would have been **Z** rather than **Y**.

Counterfactual Visual Explanations

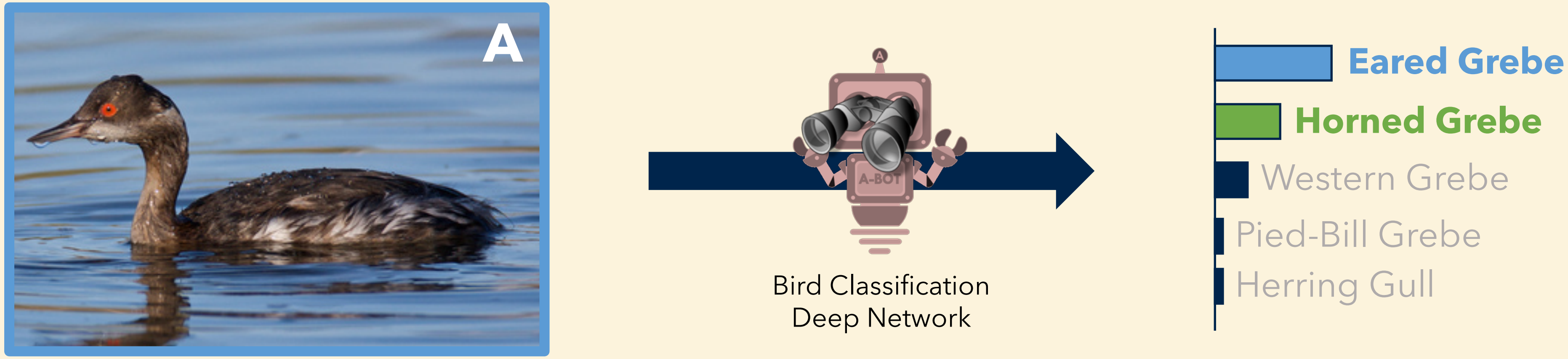


An XAI Question: Why did the model predict **Eared Grebe** instead of **Horned Grebe**?

Explanation through Counterfactual:

What would have to change in **image A** to make the model predict **Horned Grebe**?

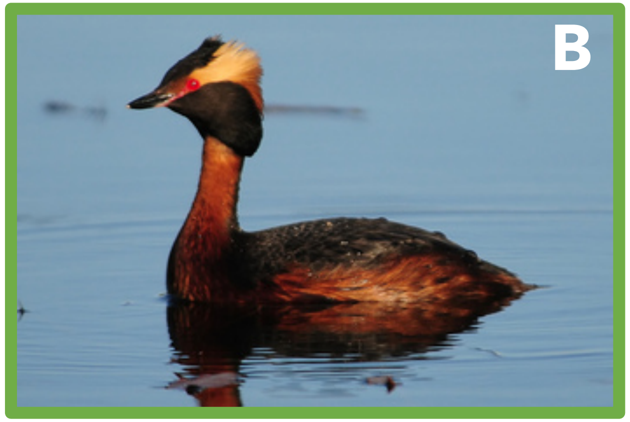
Counterfactual Visual Explanations



An XAI Question: Why did the model predict **Eared Grebe** instead of **Horned Grebe**?

Explanation through Counterfactual:

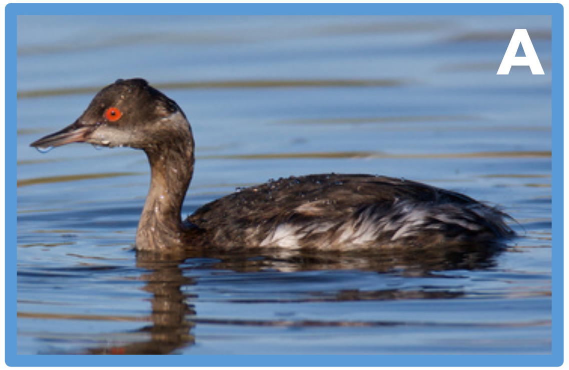
What would have to change in **image A** to make the model predict **Horned Grebe**?

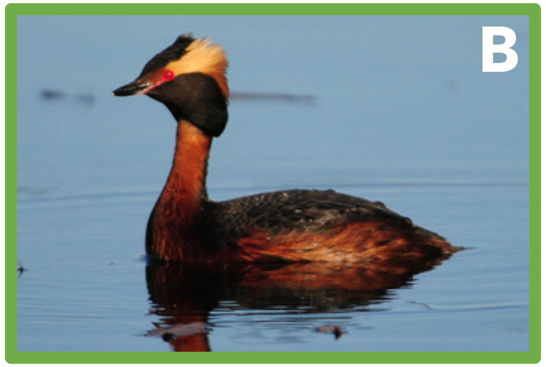


An image where the network predicts Horned Grebe.

Counterfactual Visual Explanations

What would have to change in **image A** to make the model predict **Horned Grebe**?

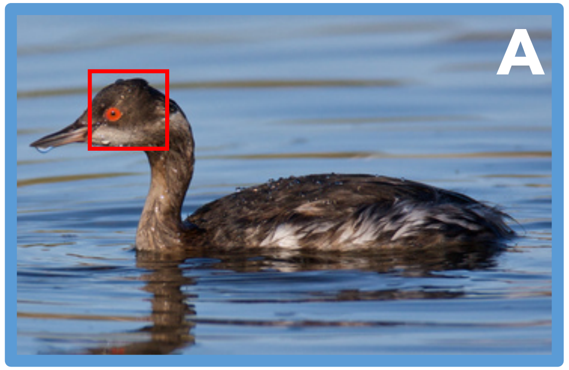


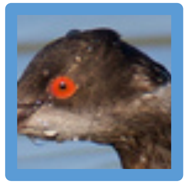
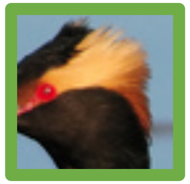


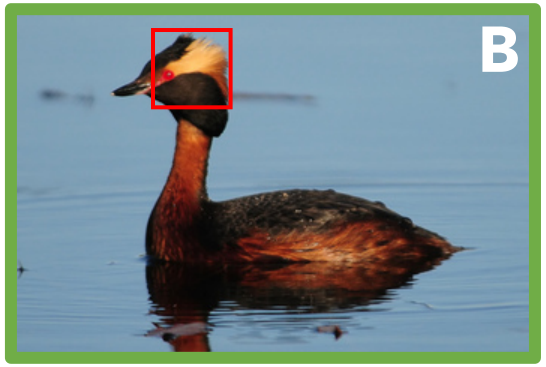
An image where the network predicts Horned Grebe.

Counterfactual Visual Explanations

What would have to change in **image A** to make the model predict **Horned Grebe**?



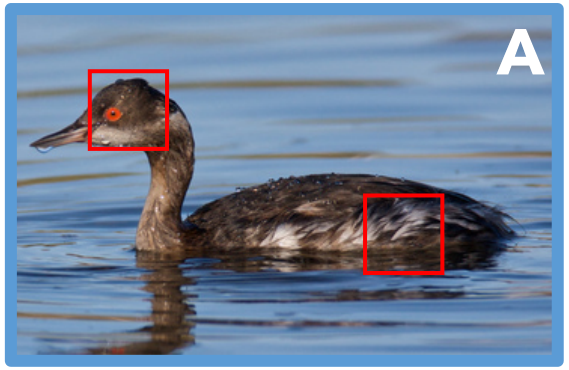
If  looked more like 

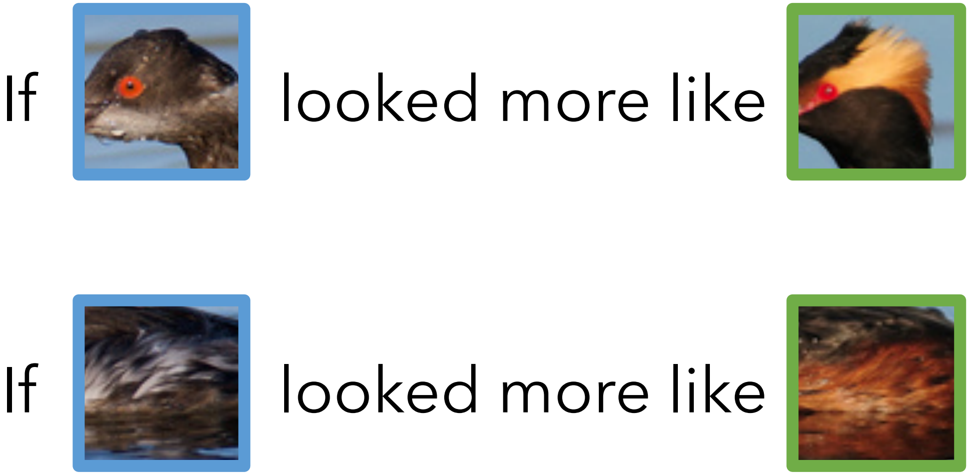


An image where the network predicts Horned Grebe.

Counterfactual Visual Explanations

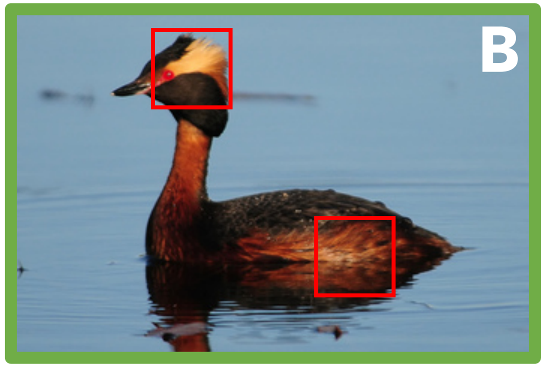
What would have to change in **image A** to make the model predict **Horned Grebe**?





If looked more like

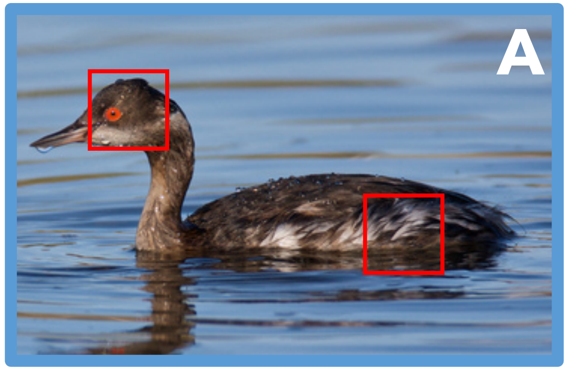
If looked more like

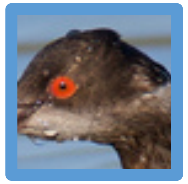


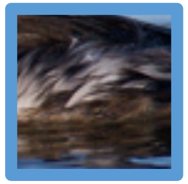
An image where the network predicts Horned Grebe.

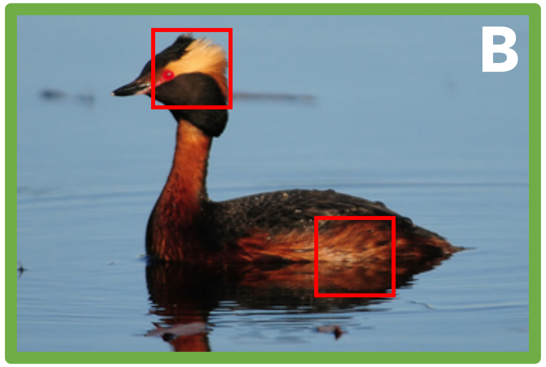
Counterfactual Visual Explanations

What would have to change in **image A** to make the model predict **Horned Grebe**?



If  looked more like

If  looked more like

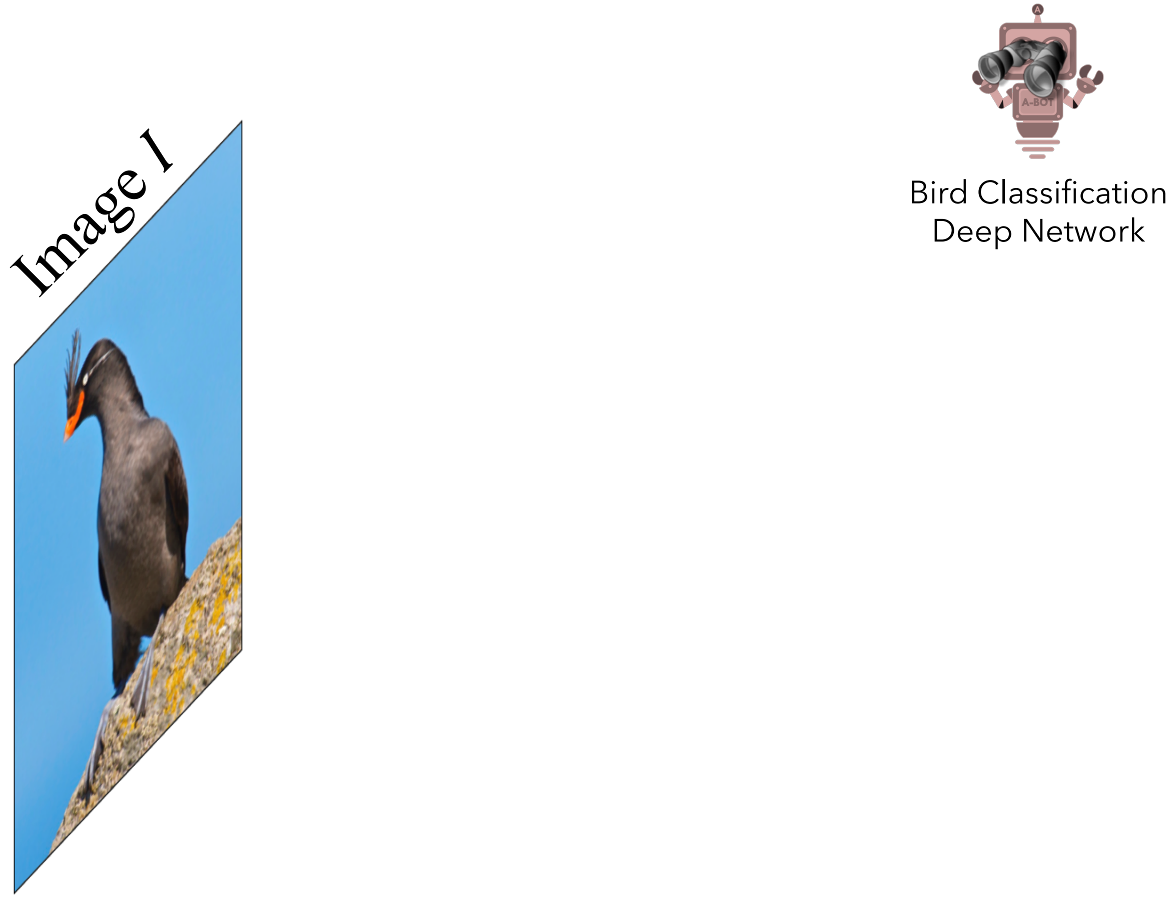


An image where the network predicts Horned Grebe.

How can we identify these region pairs important to the model?

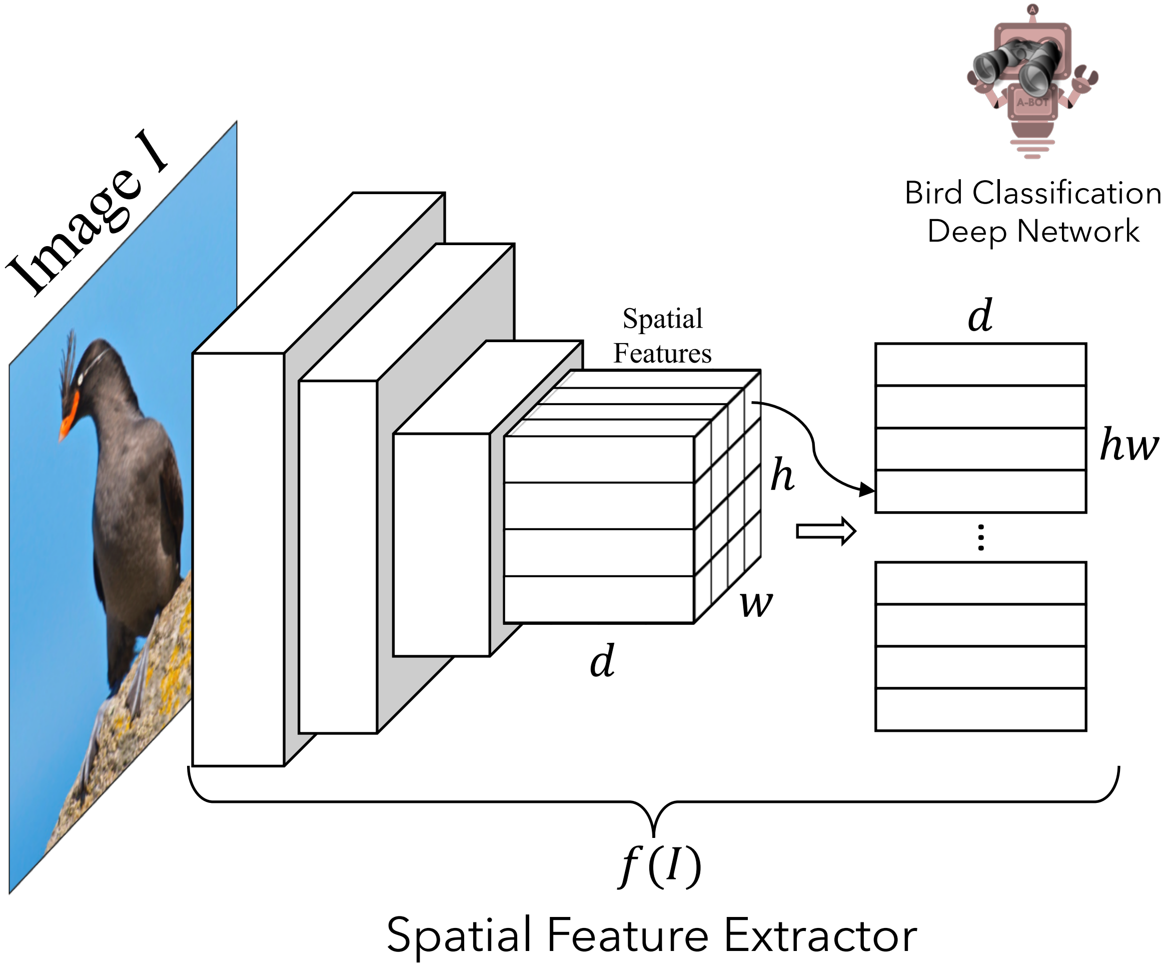
Counterfactual Visual Explanations

How can we identify these region pairs important to the model?



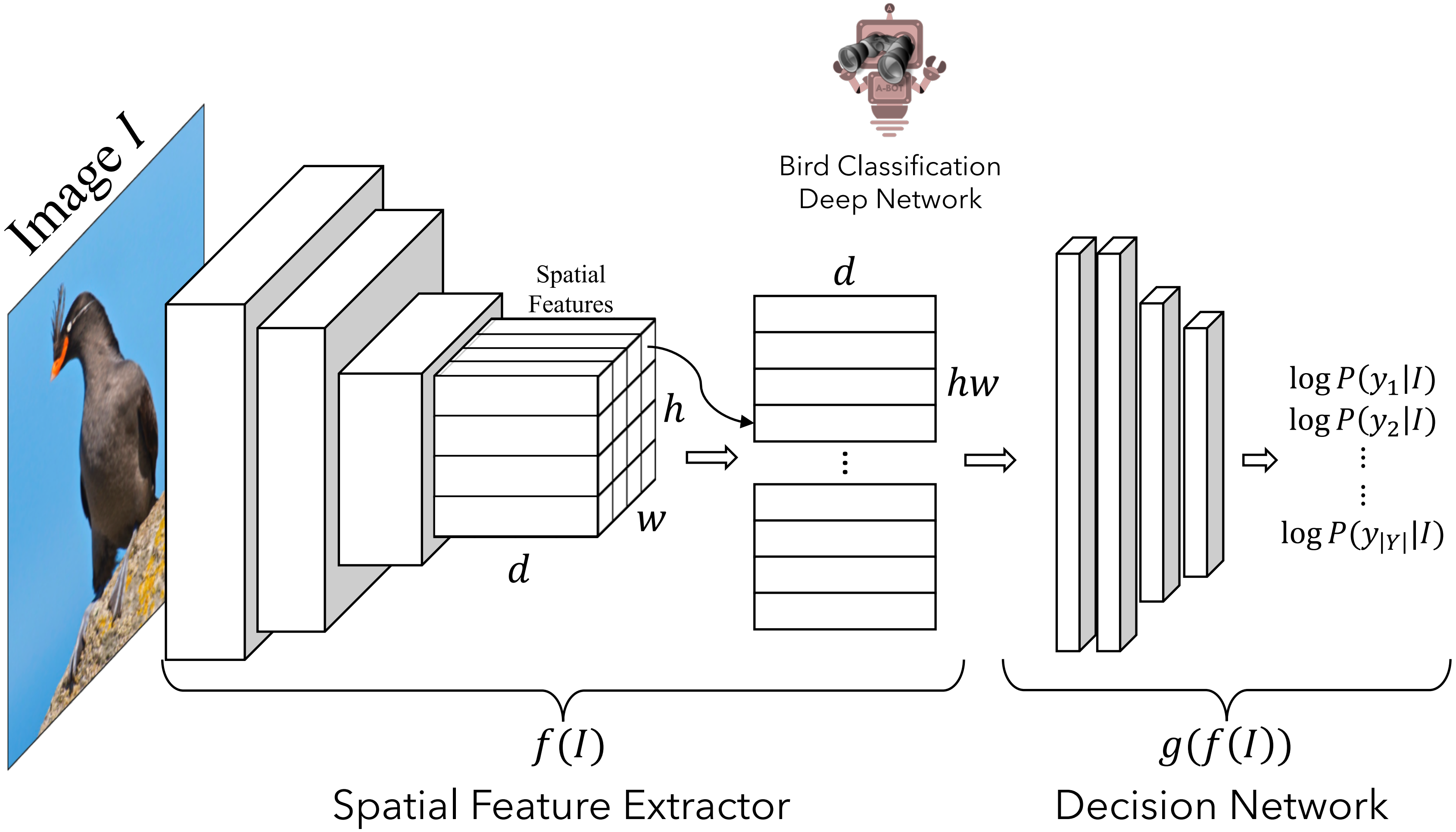
Counterfactual Visual Explanations

How can we identify these region pairs important to the model?

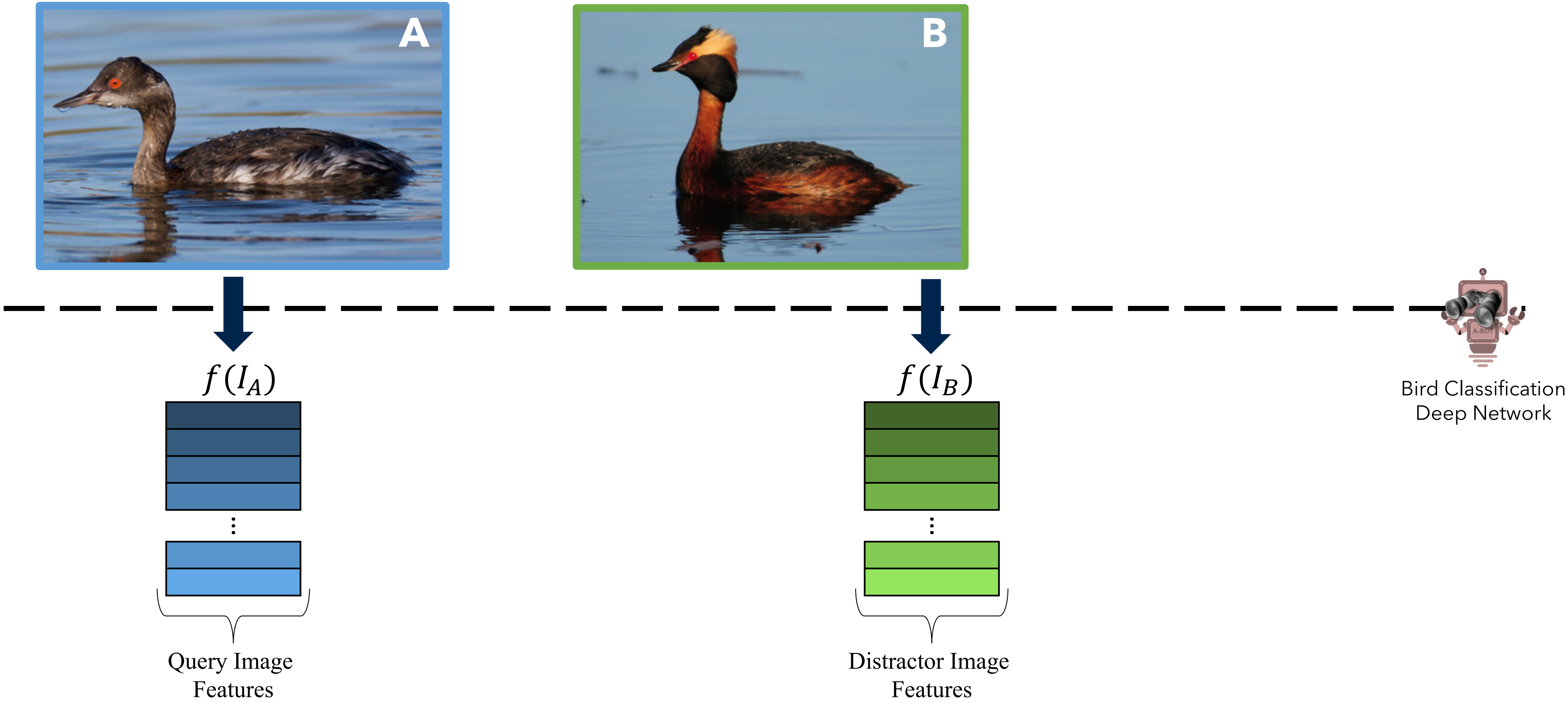


Counterfactual Visual Explanations

How can we identify these region pairs important to the model?



Counterfactual Visual Explanations



$f\,(I_A)$

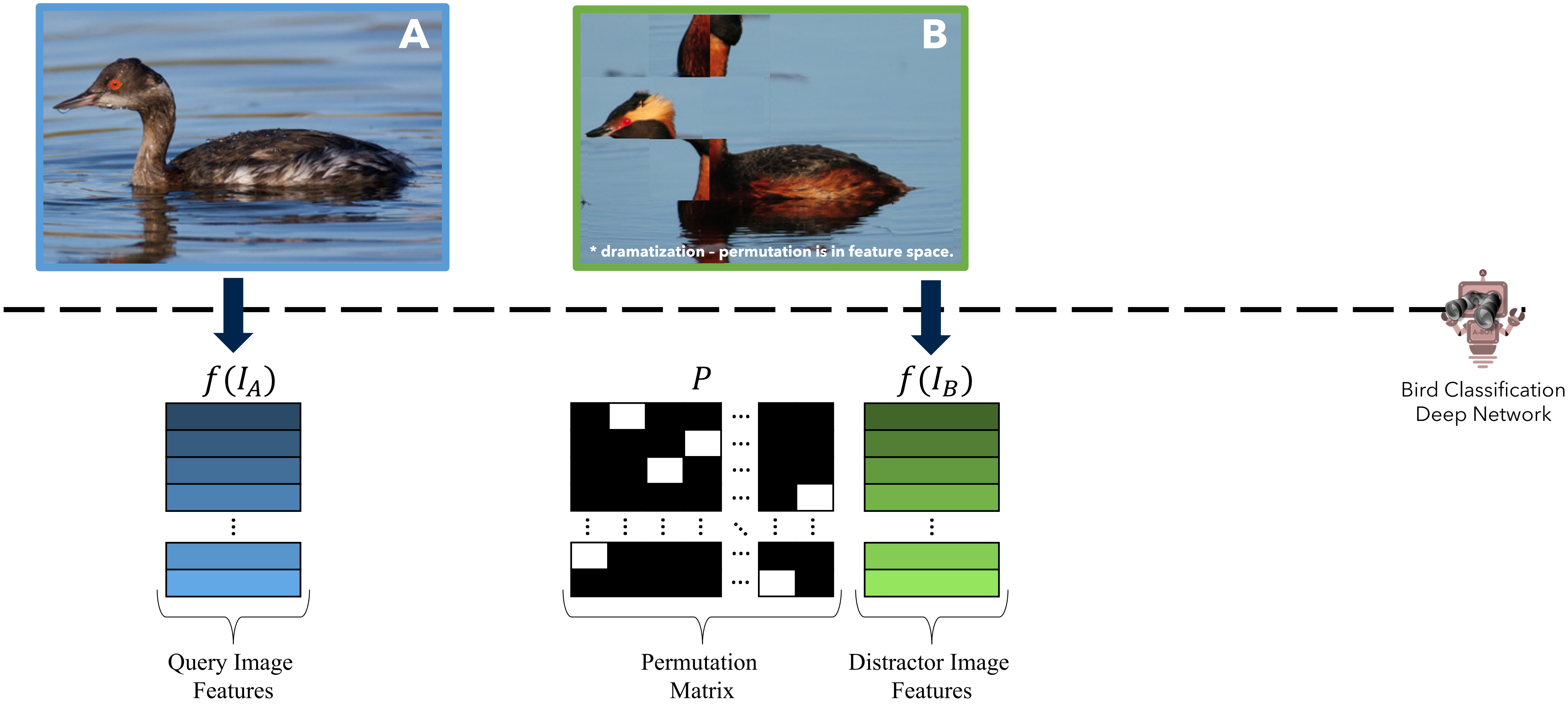
Query Image Features

$f\,(I_B)$

Distractor Image Features

Bird Classification Deep Network

Counterfactual Visual Explanations



* dramatization - permutation is in feature space.

$f(I_A)$

P

$f(I_B)$

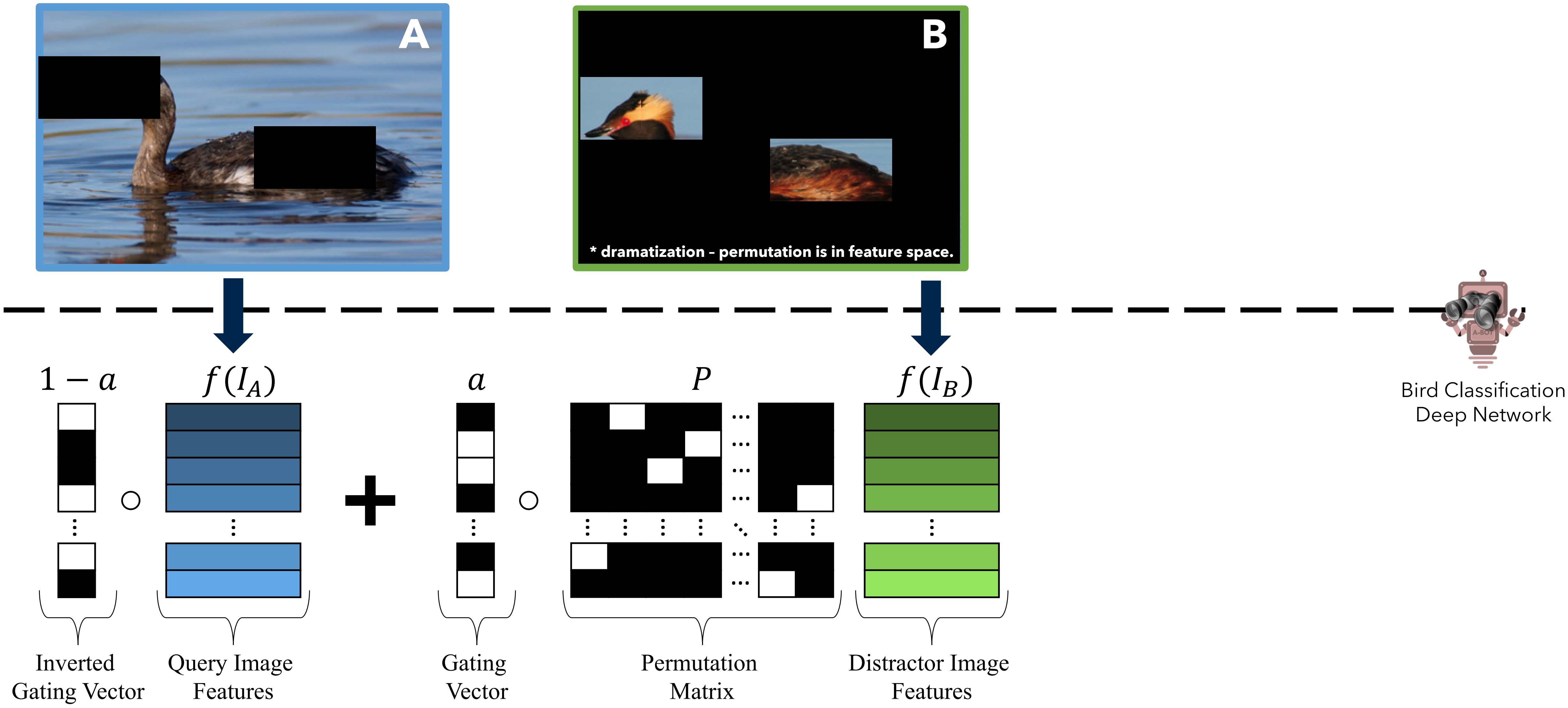
Query Image Features

Permutation Matrix

Distractor Image Features

Bird Classification Deep Network

Counterfactual Visual Explanations



* dramatization – permutation is in feature space.

$1 - a$ $\circ$ $f(I_A)$ $+$ a $\circ$ P $f(I_B)$

Inverted Gating Vector

Query Image Features

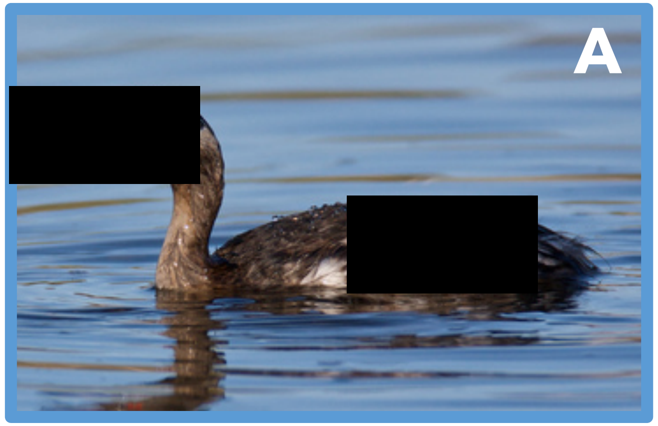
Gating Vector

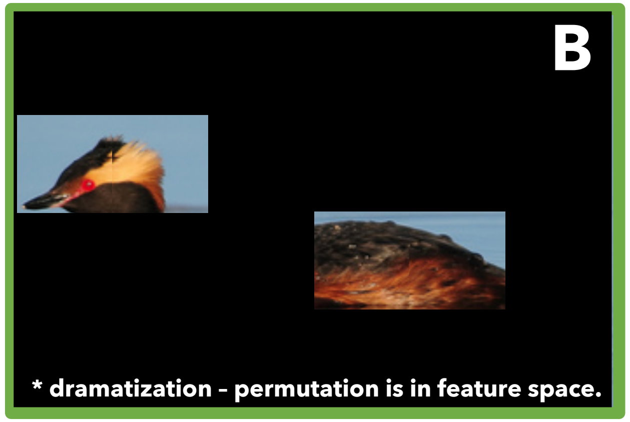
Permutation Matrix

Distractor Image Features

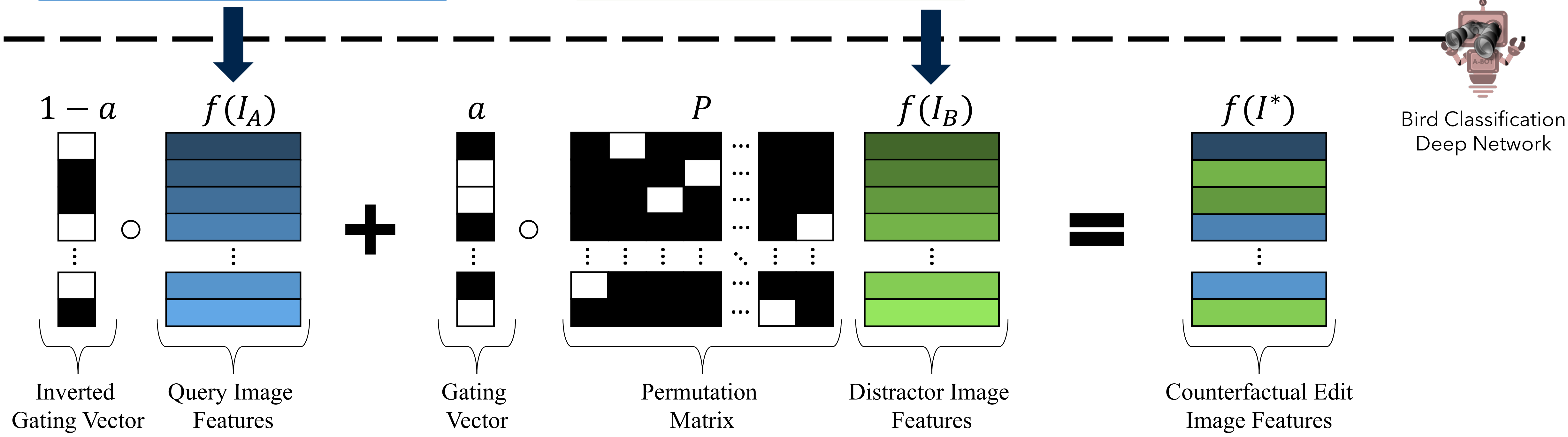
Bird Classification Deep Network

Counterfactual Visual Explanations





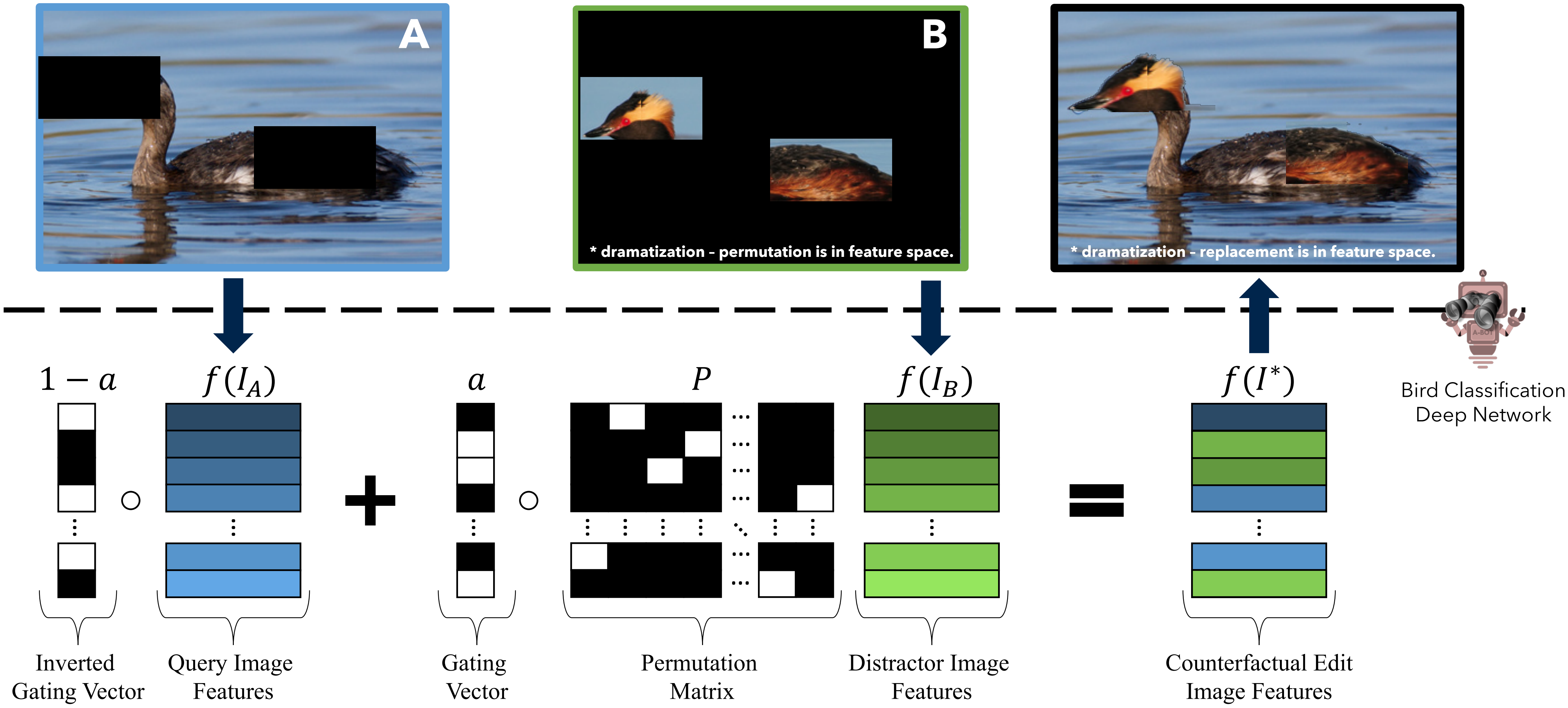
* dramatization – permutation is in feature space.



$$\underbrace{1-a}_{\text{Inverted Gating Vector}} \circ \underbrace{f(I_A)}_{\text{Query Image Features}} + \underbrace{a}_{\text{Gating Vector}} \circ \underbrace{P}_{\text{Permutation Matrix}} \; \underbrace{f(I_B)}_{\text{Distractor Image Features}} = \underbrace{f(I^*)}_{\text{Counterfactual Edit Image Features}}$$

Bird Classification Deep Network

Counterfactual Visual Explanations



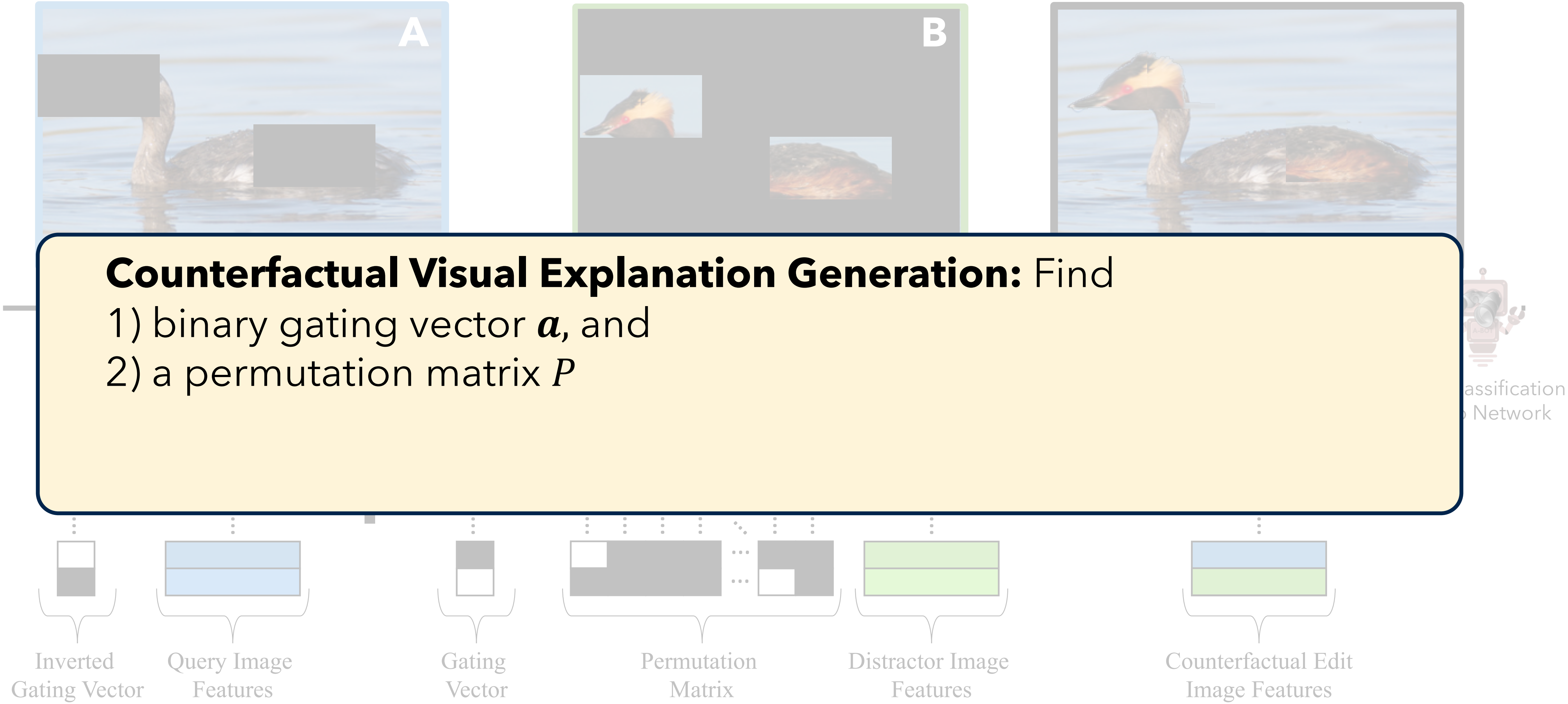
* dramatization – permutation is in feature space.

* dramatization – replacement is in feature space.

Bird Classification Deep Network

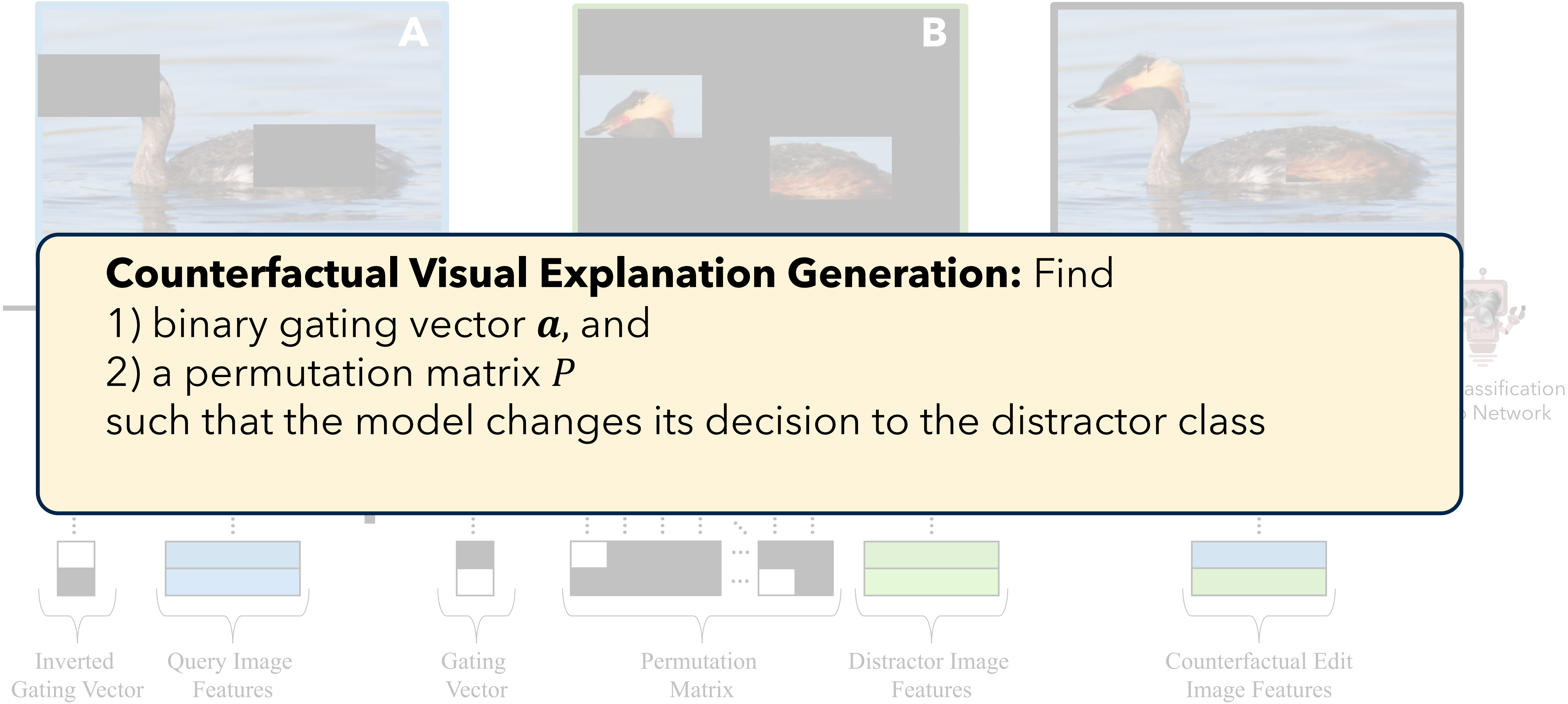
$$\underbrace{1-a}_{\text{Inverted Gating Vector}} \circ \underbrace{f(I_A)}_{\text{Query Image Features}} + \underbrace{a}_{\text{Gating Vector}} \circ \underbrace{P}_{\text{Permutation Matrix}} \underbrace{f(I_B)}_{\text{Distractor Image Features}} = \underbrace{f(I^*)}_{\text{Counterfactual Edit Image Features}}$$

Counterfactual Visual Explanations



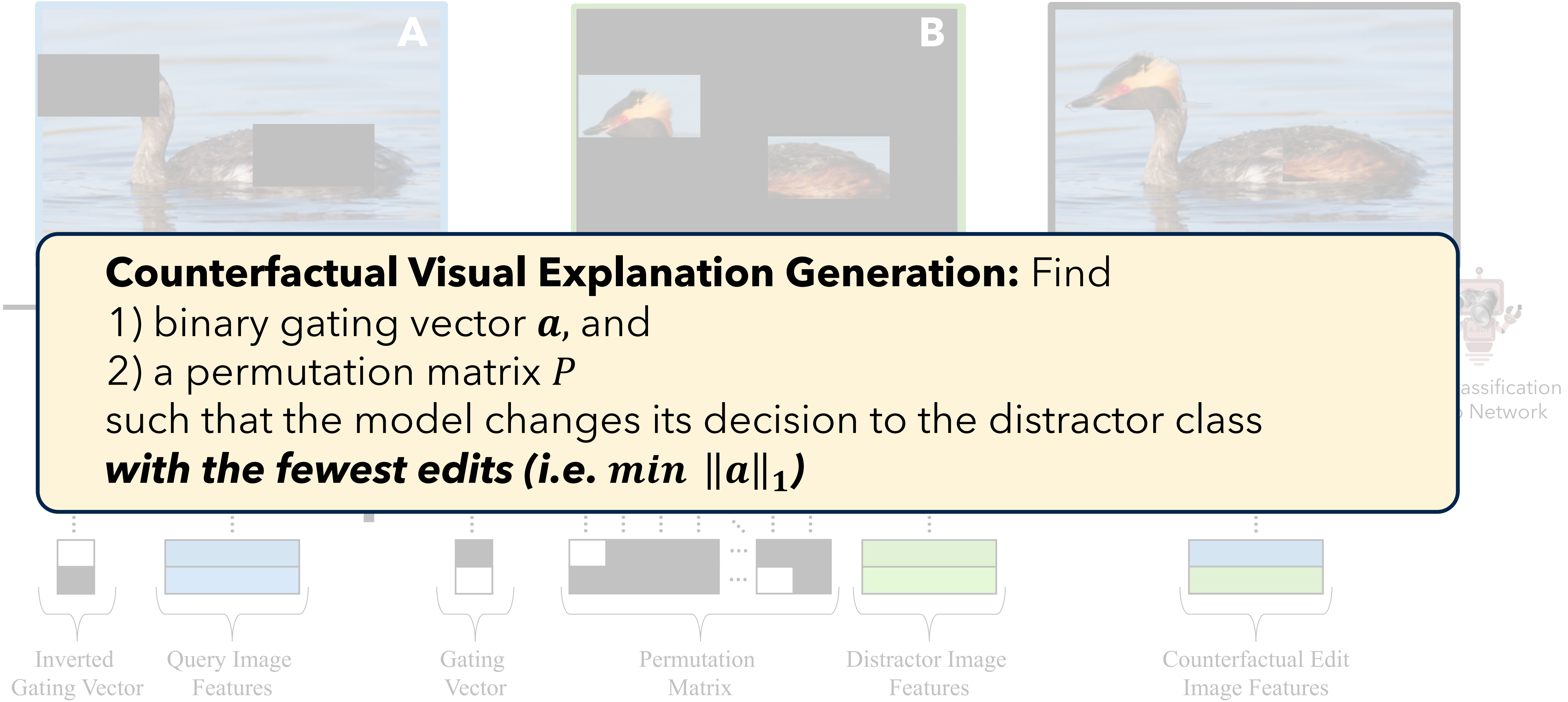
Counterfactual Visual Explanation Generation: Find
1) binary gating vector $\boldsymbol{a}$, and
2) a permutation matrix P

Counterfactual Visual Explanations



Counterfactual Visual Explanation Generation: Find
1) binary gating vector $\boldsymbol{a}$, and
2) a permutation matrix P
such that the model changes its decision to the distractor class

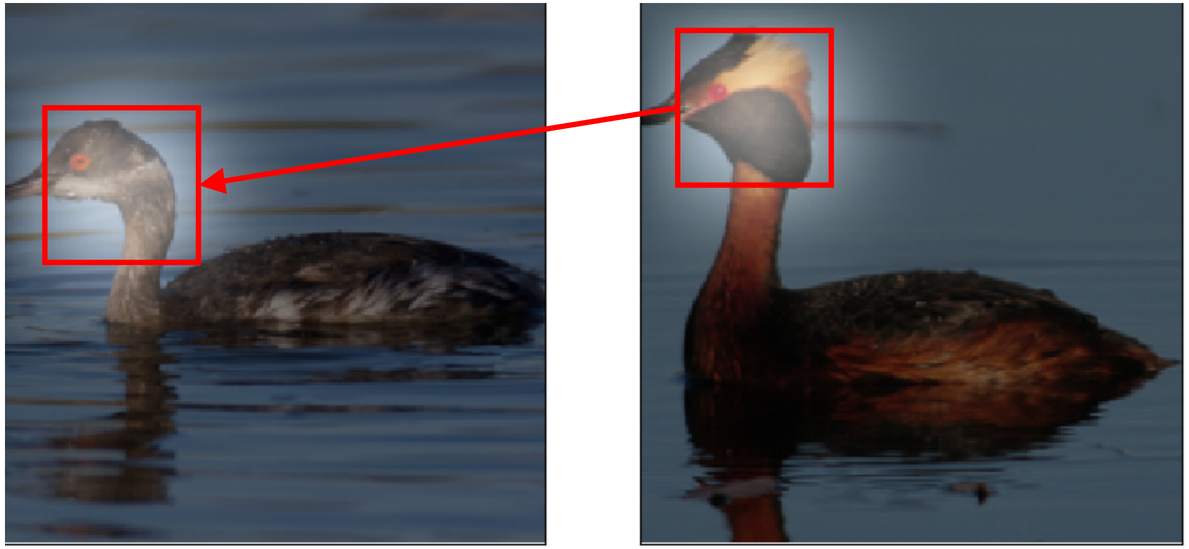
Counterfactual Visual Explanations



Counterfactual Visual Explanation Generation: Find
1) binary gating vector $\boldsymbol{a}$, and
2) a permutation matrix P
such that the model changes its decision to the distractor class
with the fewest edits (i.e. $min\ \|a\|_1$)

Results – Single Edit

If the highlighted region in **image A**
looked like the highlighted region in **image B**,
then image A is more likely to be classified as ***class B****.*



Query Image A

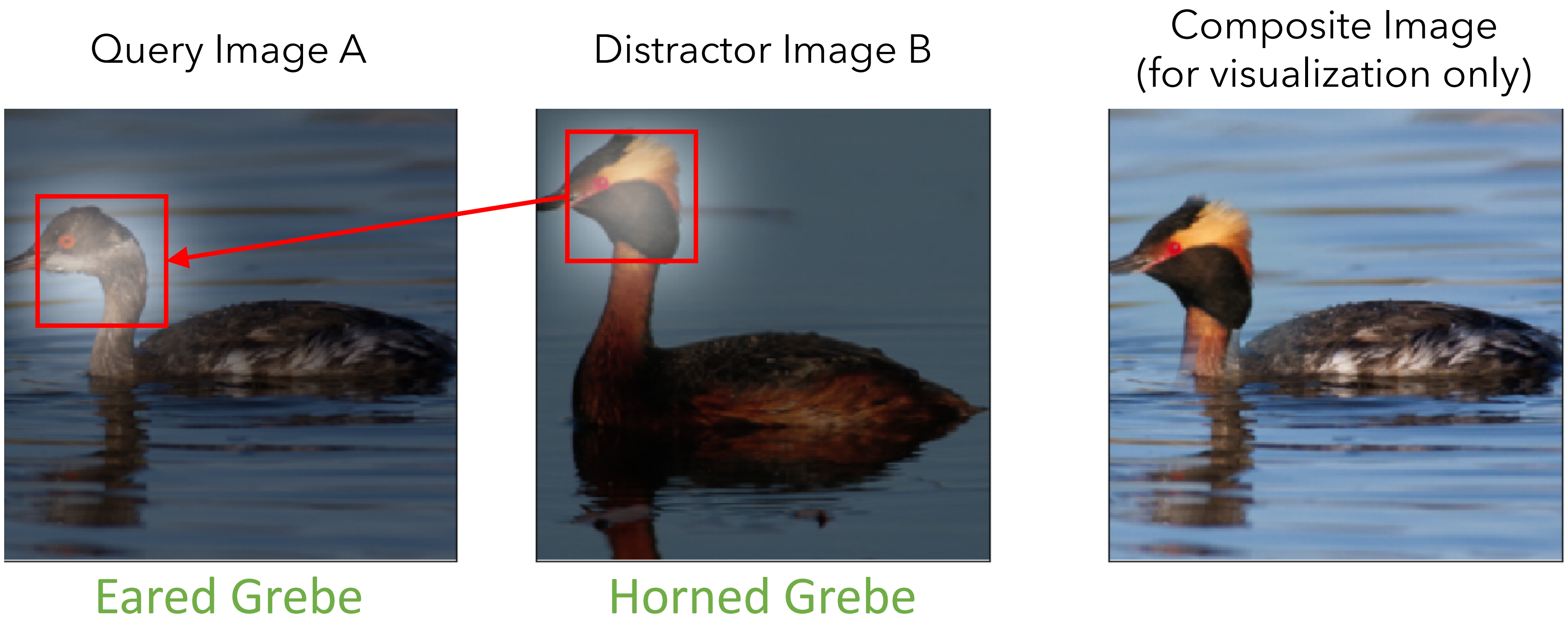
Distractor Image B

Eared Grebe

Horned Grebe

Results – Single Edit

If the highlighted region in **image A**
looked like the highlighted region in **image B**,
*then image A is more likely to be classified as **class B**.*



Query Image A

Distractor Image B

Composite Image
(for visualization only)

Eared Grebe

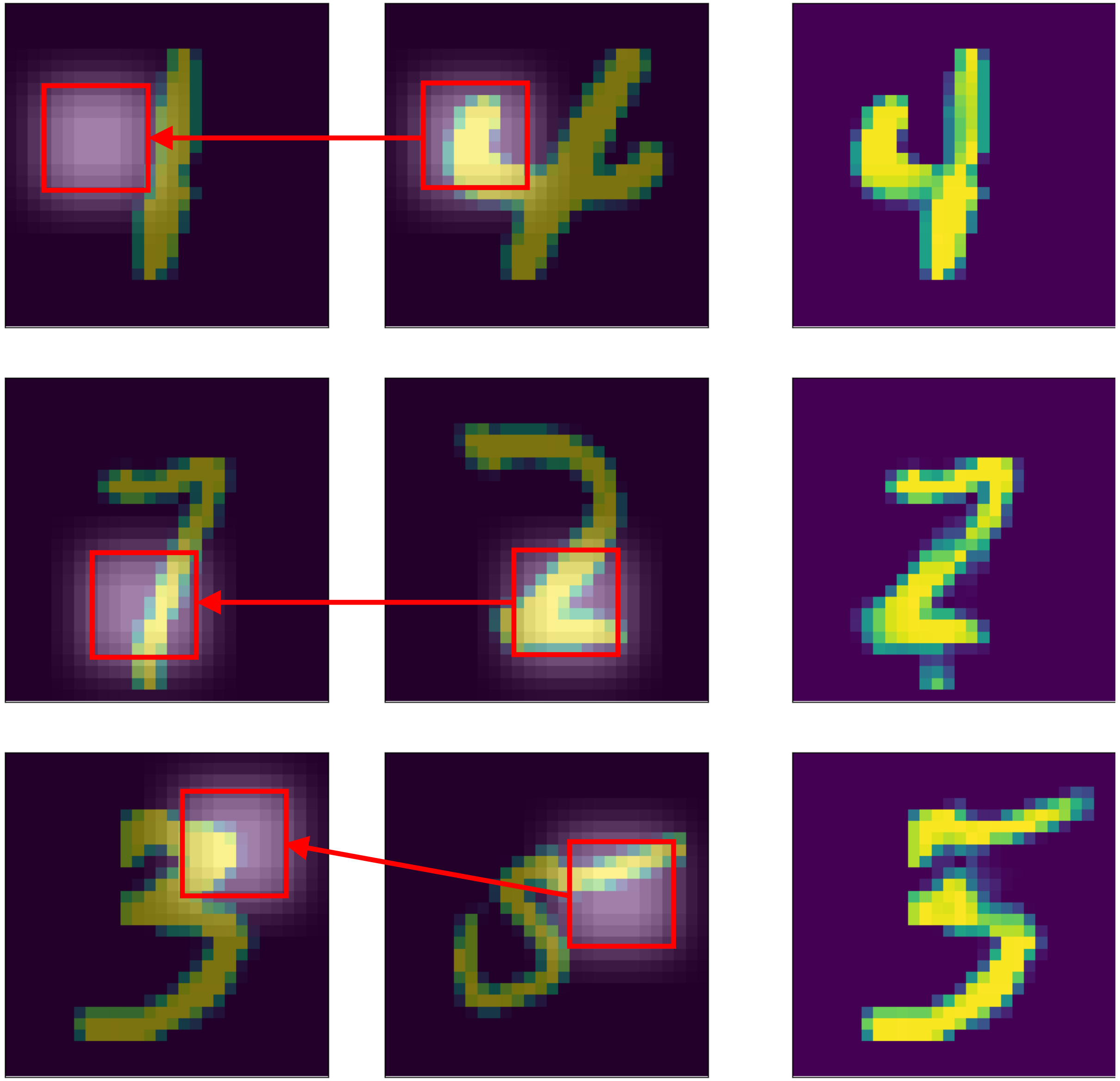
Horned Grebe

Results – Single Edit

Query Image A

Distractor Image B

Composite Image (for visualization only)

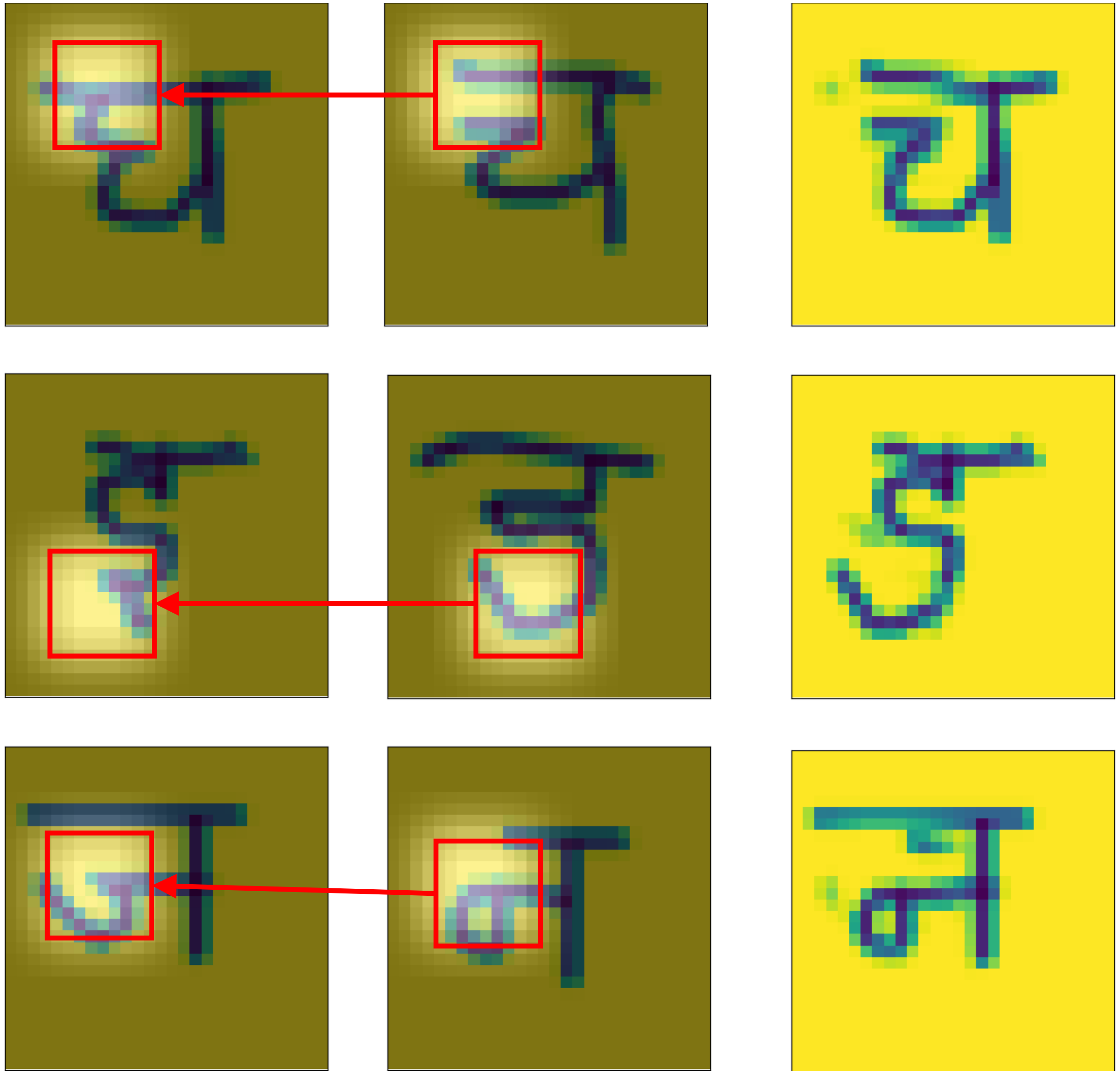


Results – Single Edit

Query Image A

Distractor Image B

Composite Image (for visualization only)

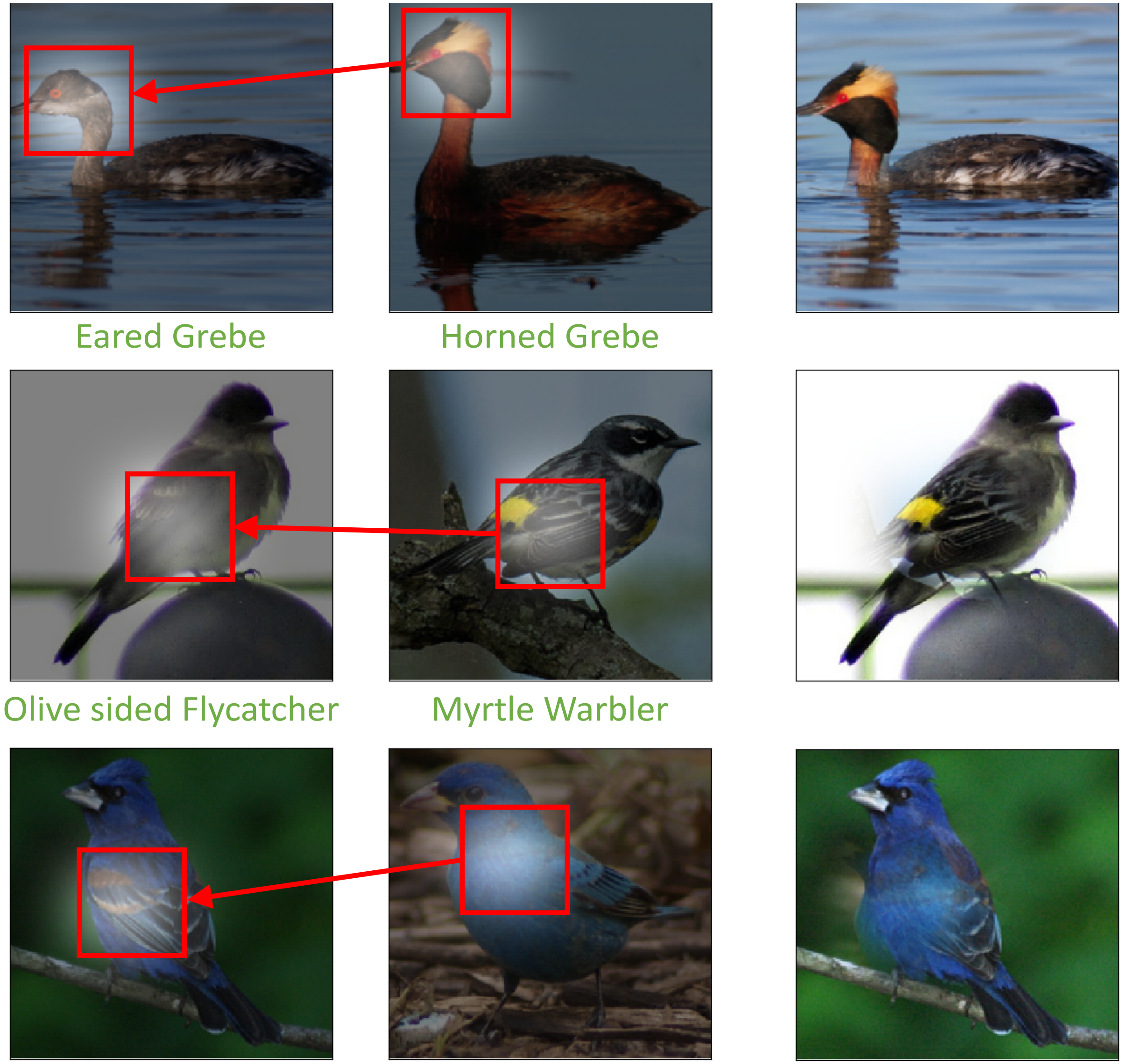


Results – Single Edit

Query Image A

Distractor Image B

Composite Image (for visualization only)



Eared Grebe

Horned Grebe

Olive sided Flycatcher

Myrtle Warbler

Blue Grosbeak

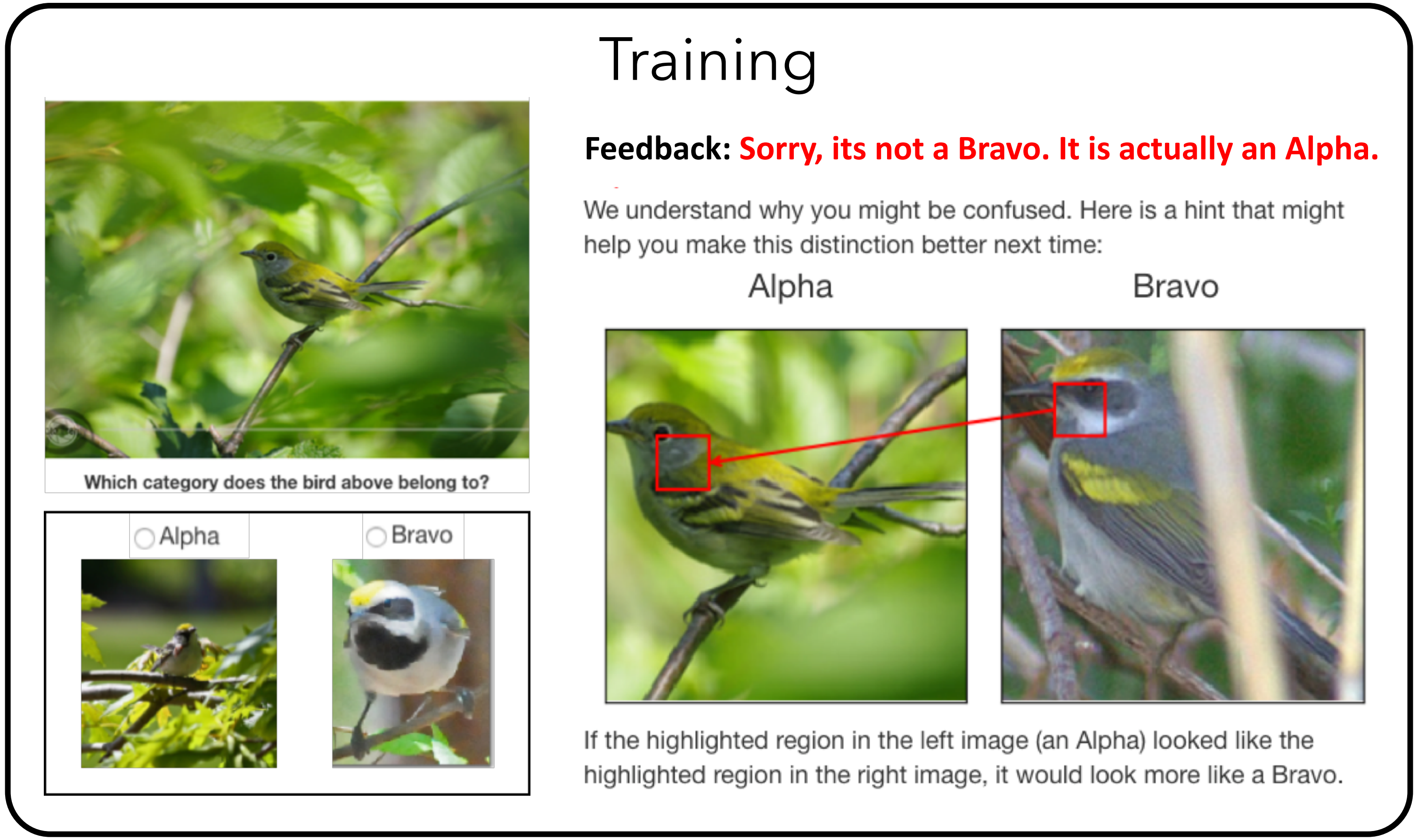
Indigo Bunting

Machine Teaching – Bird Classification

Do our **counterfactual explanations** help untrained participants learn to identify fine-grained classes?

Machine Teaching – Bird Classification

Do our **counterfactual explanations** help untrained participants learn to identify fine-grained classes?



Training

Which category does the bird above belong to?

○ Alpha ○ Bravo

Feedback: Sorry, its not a Bravo. It is actually an Alpha.

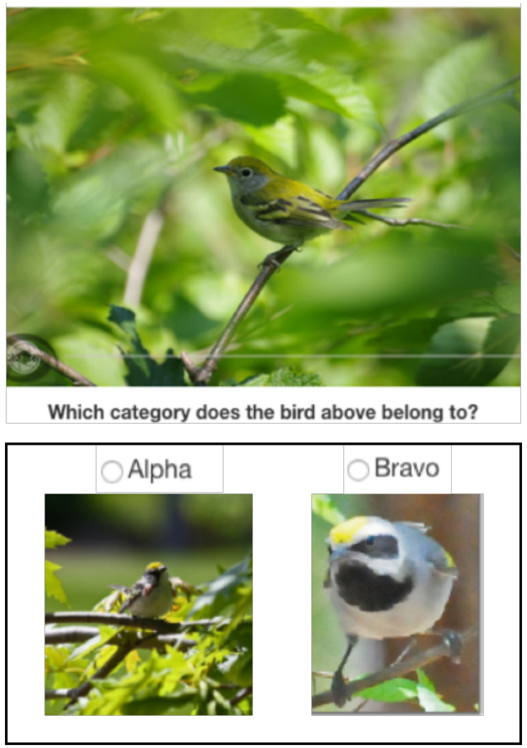
We understand why you might be confused. Here is a hint that might help you make this distinction better next time:

If the highlighted region in the left image (an Alpha) looked like the highlighted region in the right image, it would look more like a Bravo.

Machine Teaching – Bird Classification

Do our **counterfactual explanations** help untrained participants learn to identify fine-grained classes?

Training

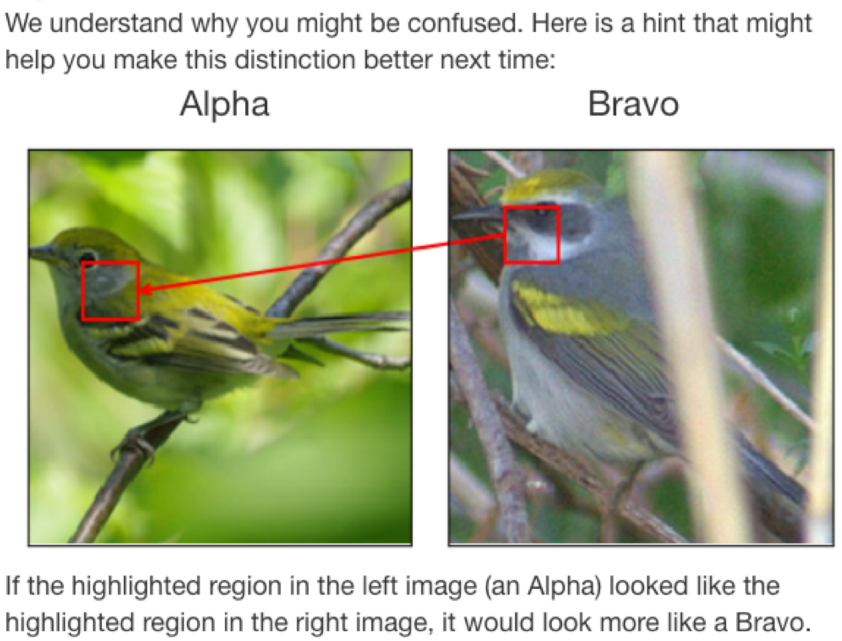


Which category does the bird above belong to?

○ Alpha

○ Bravo

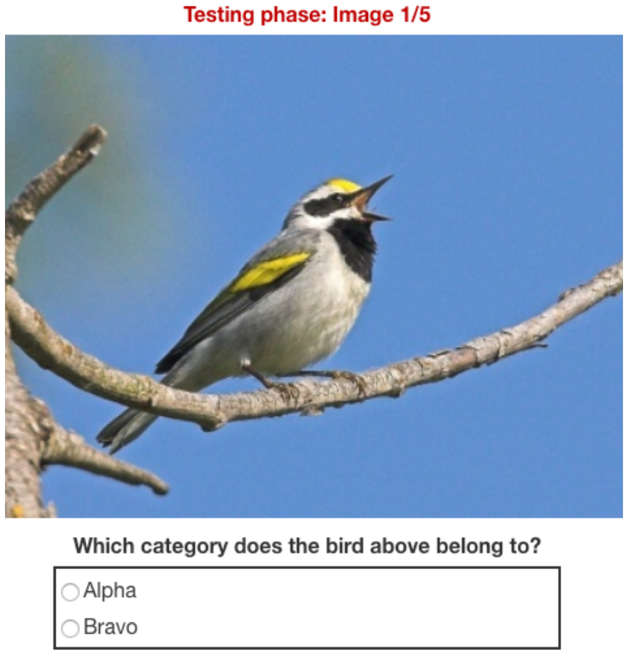
Feedback: Sorry, its not a Bravo. It is actually an Alpha.



We understand why you might be confused. Here is a hint that might help you make this distinction better next time:

If the highlighted region in the left image (an Alpha) looked like the highlighted region in the right image, it would look more like a Bravo.

Testing



Testing phase: Image 1/5

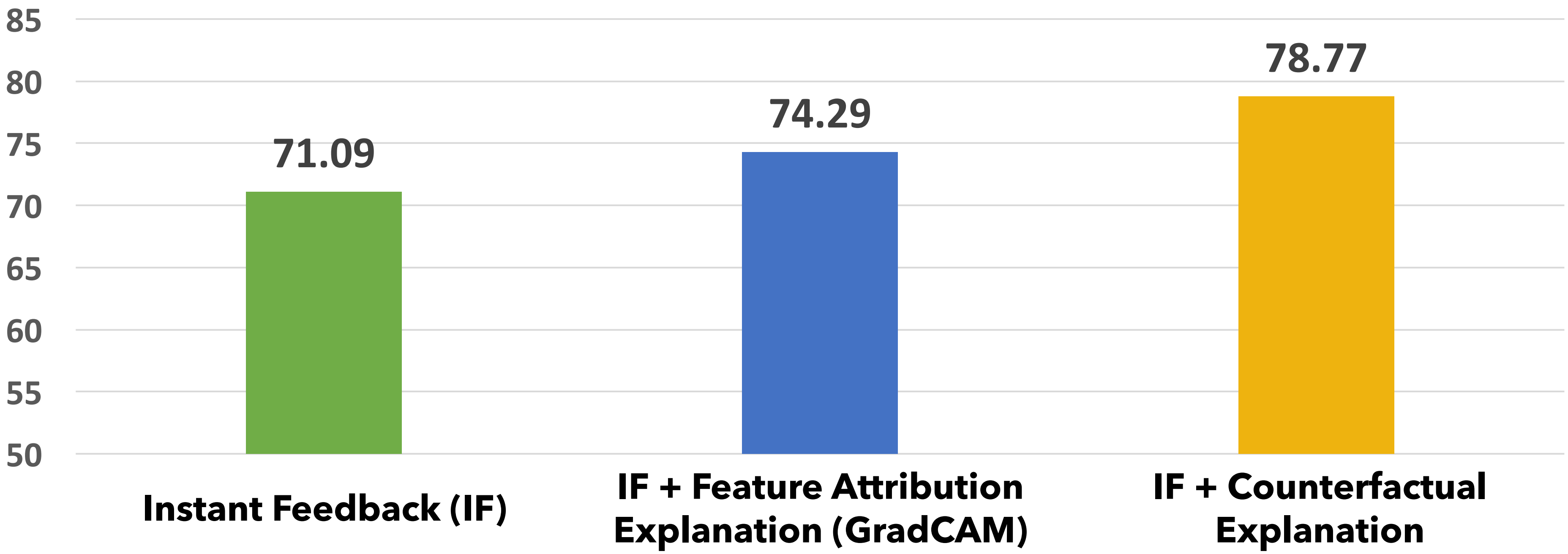
Which category does the bird above belong to?

○ Alpha

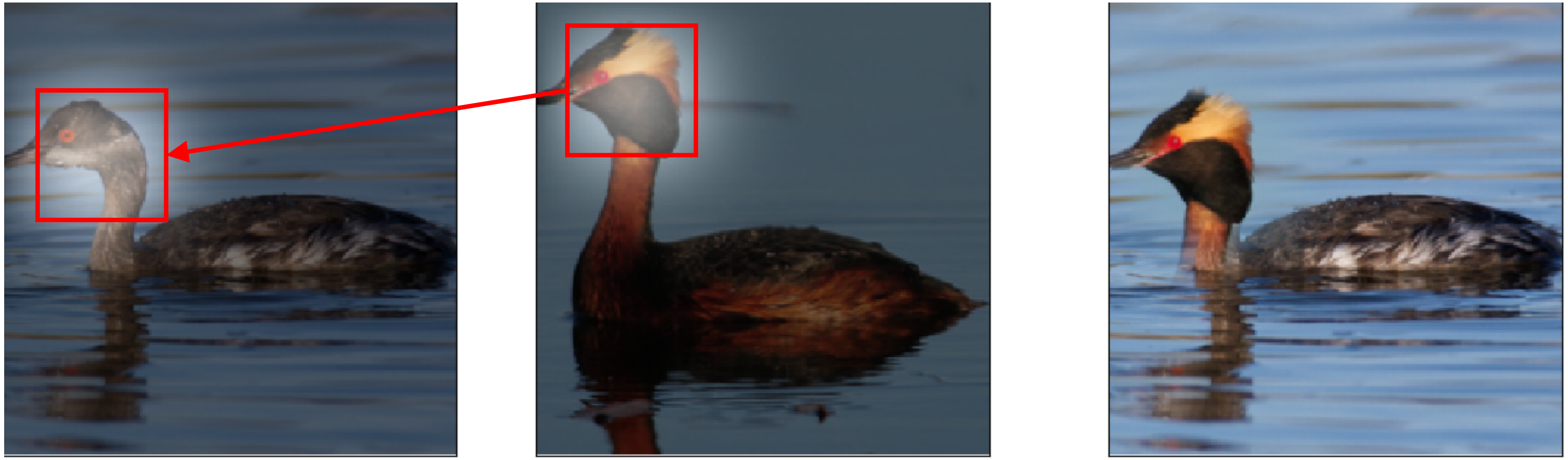
○ Bravo

Machine Teaching – Bird Classification

Do our **counterfactual explanations** help untrained participants learn to identify fine-grained classes?



Counterfactual Visual Explanations



Query Image A

Distractor Image B

Composite Image
(for visualization only)

Eared Grebe

Horned Grebe

Questions?

Stop by our poster at #149 in Pacific Ballroom!