Tackling Shortcut Learning in Deep Neural Networks: An Iterative Approach with Interpretable Models

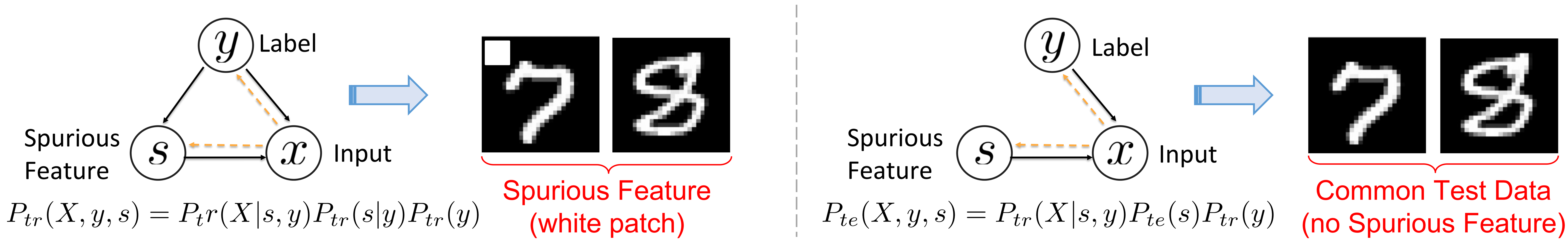
Shantanu Ghosh[1], Ke Yu[2], Forough Arabshahi[3], Kayhan Batmanghelich[1]

[1]Dept. Of Electrical and Computer Engineering, Boston University

[2]Intelligent Systems Program (ISP), University of Pittsburgh, [3]Meta AI

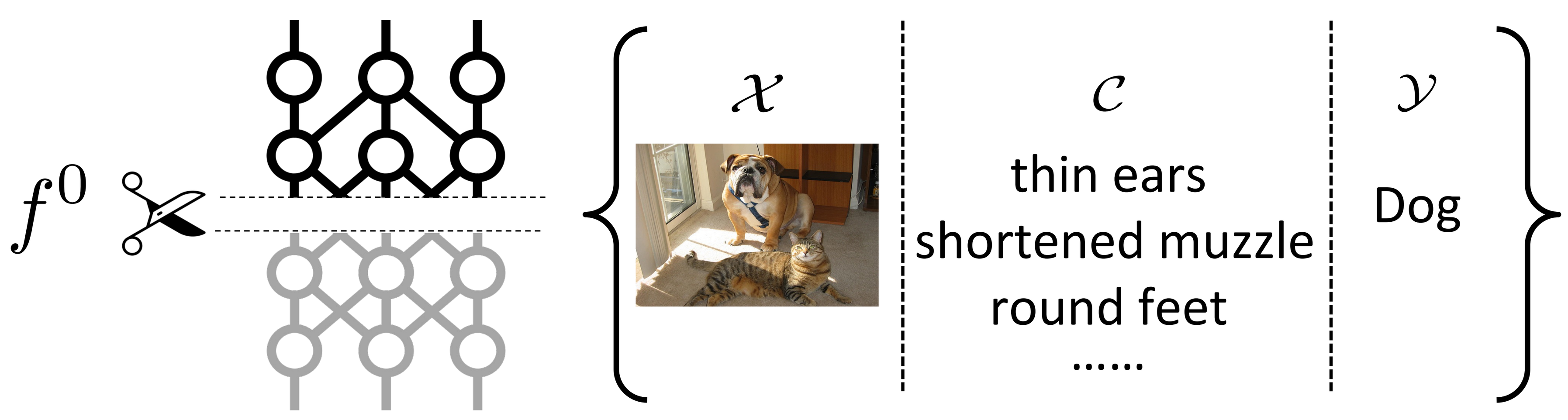
TLDR: We use concept based interpretable method to mitigate the problem of shortcut learning.

What is shortcut?

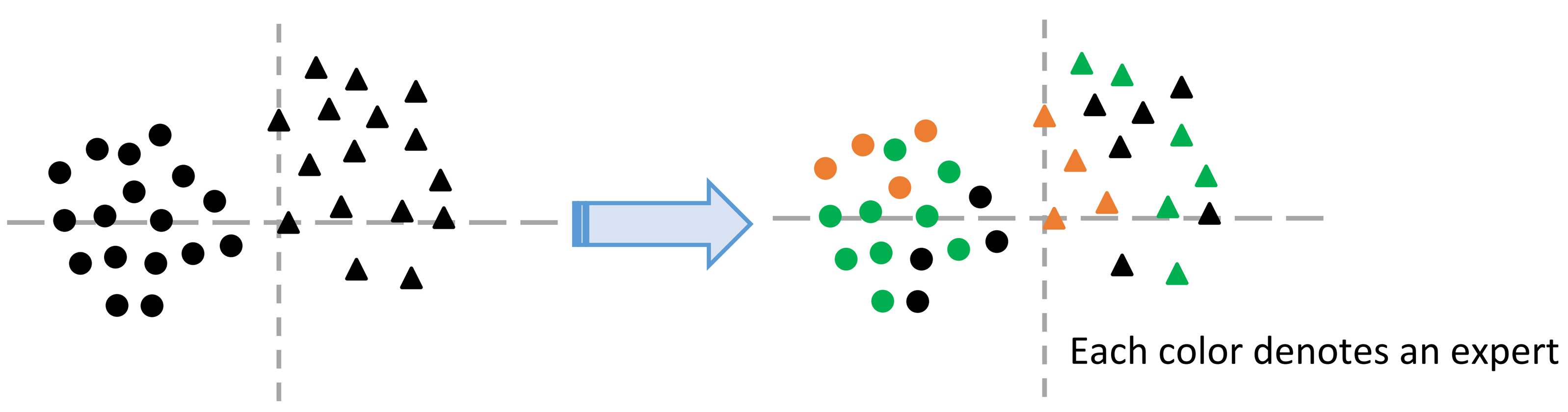


GAP: Existing shortcut elimination methods are not interpretable.

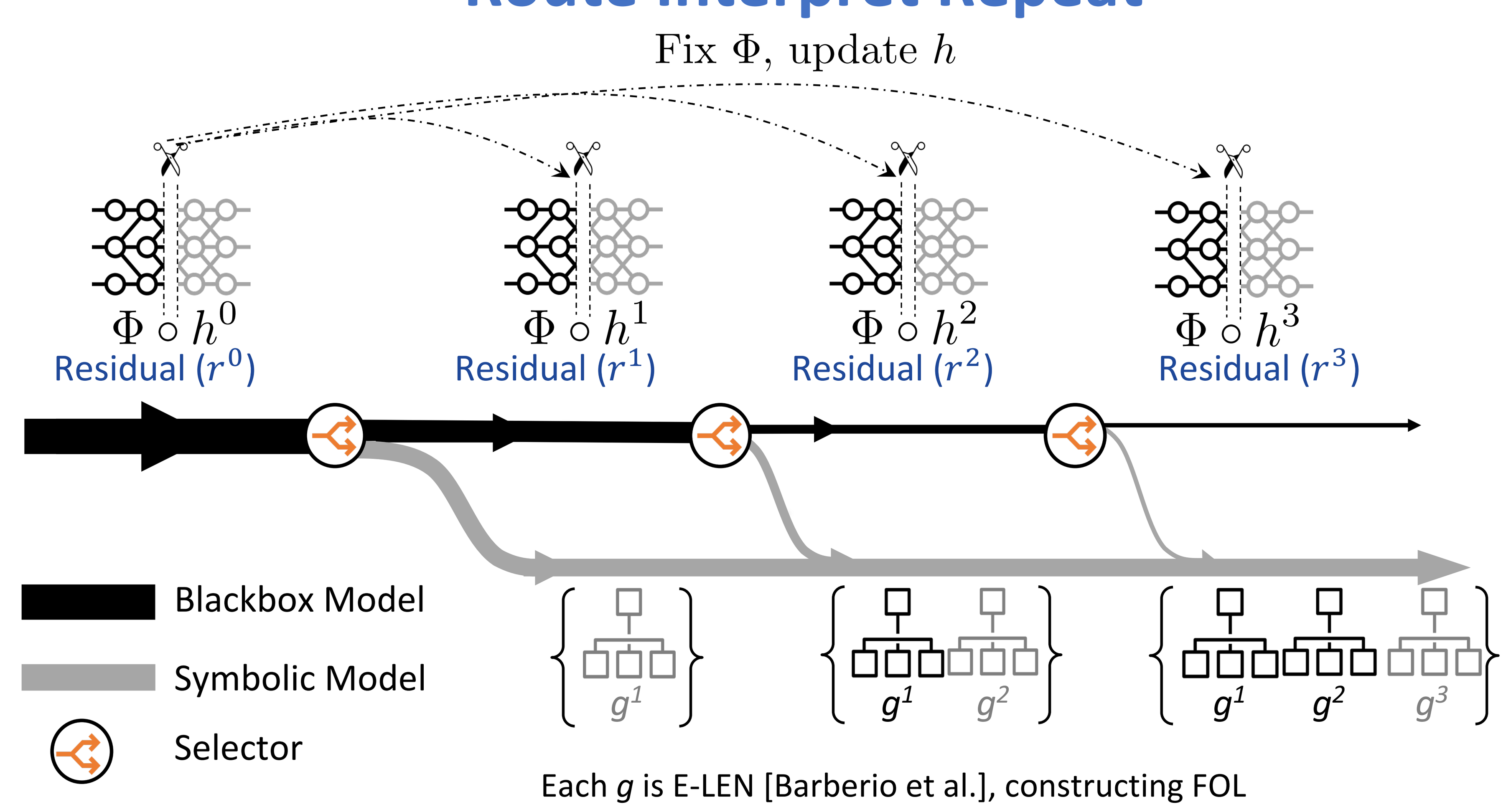
Assumption



Carve out interpretable models from Black box



Route Interpret Repeat



Each *g* is E-LEN [Barberio et al.], constructing FOL

* **SelectiveNet [Geifman et al.] optimization**
* **Continue till at least 90% samples covered**

Algorithm 1 Applying MoIE to eliminate shortcuts

1: **Input:** $\mathcal{D} = \{x_j, c_j, y_j\}_{j=1}^n$; biased BB $f^0 = h^0(\Phi(.))$; The total iterations K; Coverages $\tau_1, \dots, \tau_K$. Freeze Φ.

2: Using (Yuksekgonul et al., 2022) learn the projection t to predict the concept value.

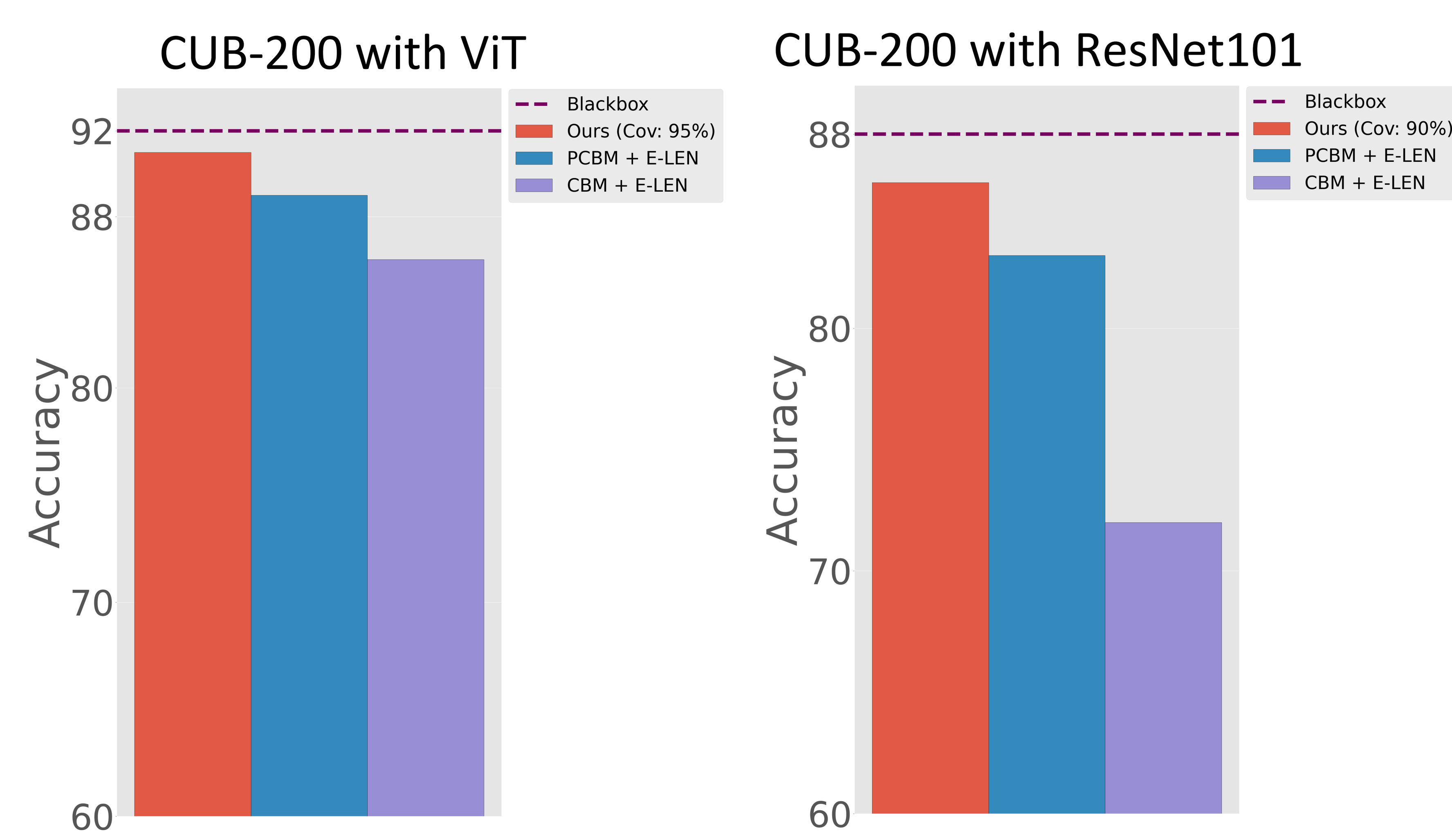
3: **Detection step.** Learn the experts in MoIE $\{g\}_{k=1}^K$ and extract the FOLs. The FOL contains shortcuts.

4: **Elimination step.** Consider the detected shortcut concept in the "Detection" step as metadata and finetune BB (f^0) with MDN (Lu et al., 2021) to remove the role of that shortcut.

5: Retrain t with Φ of finetuned BB to get the concepts.

6: **Verification step.** Learn MoIE $\{g\}_{k=1}^K$ again from retrained t and recompute the FOLs. The final FOLs do not contain spurious concepts as they have been eliminated in the "Elimination step".

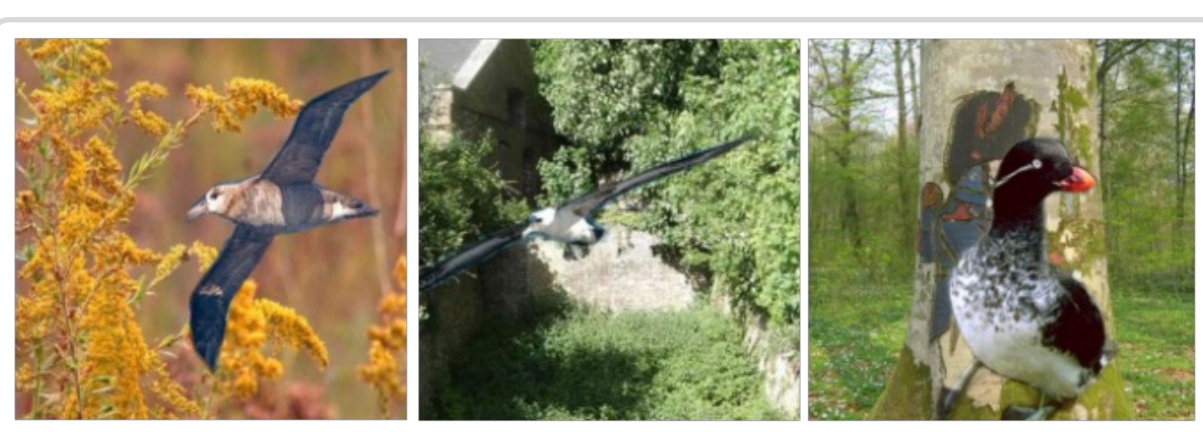
Not compromising the accuracy



Comparison w/ other methods on Waterbirds dataset

Method	Avg Acc.	Worst Acc.
ERM (Wah et al., 2011)	97.0 ± 0.2%	63.7 ± 1.9%
ERM+aug (Wah et al., 2011)	87.4 ± 0.5%	76.4 ± 2.0%
UW (Xian et al., 2018)	96.3.0 ± 0.3%	76.2 ± 1.4%
IRM (Arjovsky et al., 2020)	87.5 ± 0.7%	75.6 ± 3.1%
IB-IRM (Ahuja et al., 2022)	88.5 ± 0.9%	76.5 ± 1.2%
V-REx (Krueger et al., 2021)	88.0 ± 1.4%	73.6 ± 0.2%
CORAL (Sun & Saenko, 2016)	90.3 ± 1.1%	79.8 ± 1.8%
Fish (Shi et al., 2021)	85.6 ± 0.4%	64.0 ± 0.3%
GroupDRO (Sagawa et al., 2019)	91.8 ± 0.3%	90.6 ± 1.1%
JTT (Liu et al., 2021)	93.3 ± 0.3%	86.7 ± 1.5%
DM-ADA (Xu et al., 2020)	76.4 ± 0.3%	53.0 ± 1.3%
LISA (Yao et al., 2022)	91.8 ± 0.3%	88.5 ± 0.8%
BB w MDN (**ours**)	**95.01 ± 0.5%**	**94.4 ± 0.5%**
MoIE from BB w MDN (**ours**) (COVERAGE)	91.0 ± 0.5% (0.91)	93.7 ± 0.4% (0.87)
MoIE+R from BB w MDN (**ours**)	90.2 ± 0.5%	92.1 ± 0.4%

Qualitative result on Waterbirds dataset



Biased Blackbox

Groundtruth: WaterBird
Prediction : LandBird
Explanation : LandBird $\leftrightarrow$ WingShapeRoundedwings $\wedge$ Forest

Robust Blackbox

Groundtruth: WaterBird
Prediction : WaterBird
Explanation : WaterBird $\leftrightarrow$ BillLengthAboutTheSameAsHead $\wedge \neg$BillLengthShorterThanHead $\wedge \neg$SizeSmall5_9in $\wedge \neg$ShapePerchingLike $\wedge$ CrownColorWhite