Supplementary materials for SCOUT: Self-aware Discriminant Counterfactual Explanation

Pei Wang Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego
{pew062,nuno}@ucsd.edu

1. Comparison to attributive explanations on segmentation datasets

In the paper, we mainly showed the results on CUB200 [7] due to limited space. The results on ADE20K [8] are shown here in Table 1. The same conclusions as those in the paper can be obtained.

2. More visualization comparison to state of the art

Please see Figure 1.

3. More Visualizations of SCOUT

Please see Figure 2 on CUB200 and Figure 3 on ADE20K.

4. Implementation details

Both datasets were subject to standard normalizations. Training images were first resized to 224×224 and then randomly flipped, whereas test images were first resized to 256×256 and then center-cropped to 224×224. All images were also first converted to [0.0, 1.0] from [0, 255] and then normalized by subtracting the mean ($[0.471, 0.460, 0.454]$) and dividing by the standard deviation ($[0.267, 0.266, 0.271]$) of each RGB color channel. All results are presented on the standard CUB200 test set and the official validation set of ADE20K. Experiments were ran three times. Used classifiers and predictors are trained by standard strategies [3, 1, 2, 6, 4].

5. Attribute Assignment

The parts and attributes of the CUB200 dataset [7] are listed in Table 2 following [5].

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[4] Pei Wang and Nuno Vasconcelos. Towards realistic predictors. In *The European Conference on Computer Vision*, 2018.

[5] Pei Wang and Nuno Vasconcelos. Deliberative explanations: visualizing network insecurities. In *Advances in Neural Information Processing Systems 32*, pages 1374–1385. 2019.

[6] Xin Wang, Yujia Luo, Daniel Crankshaw, Alexey Tumanov, Fisher Yu, and Joseph E Gonzalez. Idk cascades: Fast deep learning by learning not to overthink. *arXiv preprint arXiv:1706.00885*, 2017.

[7] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[8] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Beginners						
Explanation maps	10%	20%	30%	40%	50%	Avg.
$\mathbf{a}(h_{y^*}(\mathbf{x}))$	8.31(0.02)	15.41(0.01)	21.75(0.02)	27.64(0.03)	33.19(0.04)	21.25(0.02)
$\mathbf{a}(h_{y^*}(\mathbf{x})) \cdot \mathbf{a}(h_{y^c}(\mathbf{x}))$	**8.39(0.04)**	**15.43(0.08)**	21.79(0.11)	27.70(0.13)	33.28(0.15)	21.32(0.10)
$\mathbf{a}(h_{y^*}(\mathbf{x})) \cdot \mathbf{a}(h_{y^c}(\mathbf{x})) \cdot \mathbf{a}(s^s(\mathbf{x}))$	8.30(0.05)	15.40(0.06)	21.82(0.09)	**27.81(0.11)**	**33.45(0.15)**	**21.36(0.09)**
$\mathbf{a}(h_{y^*}(\mathbf{x})) \cdot \mathbf{a}(h_{y^c}(\mathbf{x})) \cdot \mathbf{a}(s^c(\mathbf{x}))$	8.31(0.04)	15.39(0.06)	**21.83(0.09)**	**27.81(0.12)**	**33.45(0.14)**	**21.36(0.09)**
$\mathbf{a}(h_{y^*}(\mathbf{x})) \cdot \mathbf{a}(h_{y^c}(\mathbf{x})) \cdot \mathbf{a}(s^e(\mathbf{x}))$	8.35(0.02)	15.42(0.00)	21.82(0.02)	27.78(0.02)	33.38(0.03)	21.35(0.01)
Advanced users						
Explanation maps	10%	20%	30%	40%	50%	Avg.
$\mathbf{a}(h_{y^*}(\mathbf{x}))$	5.56(0.03)	8.89(0.11)	11.36(0.11)	13.32(0.21)	14.98(0.24)	10.82(0.14)
$\mathbf{a}(h_{y^*}(\mathbf{x})) \cdot \mathbf{a}(h_{y^c}(\mathbf{x}))$	5.54(0.18)	8.95(0.32)	11.55(0.41)	13.63(0.45)	15.35(0.54)	11.00(0.38)
$\mathbf{a}(h_{y^*}(\mathbf{x})) \cdot \mathbf{a}(h_{y^c}(\mathbf{x})) \cdot \mathbf{a}(s^s(\mathbf{x}))$	**5.60(0.12)**	9.04(0.25)	11.72(0.32)	13.82(0.42)	15.57(0.49)	11.15(0.32)
$\mathbf{a}(h_{y^*}(\mathbf{x})) \cdot \mathbf{a}(h_{y^c}(\mathbf{x})) \cdot \mathbf{a}(s^c(\mathbf{x}))$	5.57(0.11)	9.05(0.26)	11.72(0.34)	13.83(0.44)	15.56(0.49)	11.15(0.33)
$\mathbf{a}(h_{y^*}(\mathbf{x})) \cdot \mathbf{a}(h_{y^c}(\mathbf{x})) \cdot \mathbf{a}(s^e(\mathbf{x}))$	5.57(0.15)	**9.08(0.22)**	**11.73(0.35)**	**14.01(0.50)**	**15.62(0.58)**	**11.20(0.36)**

Table 1: Comparison to attributive explanations (ADE20K): Upper: on beginners, lower: on advanced users.

Parts	Attributes
back	back color, back pattern
beak	bill shape, bill length, bill color
belly	belly color, belly pattern
breast	breast color, breast pattern
crown	crown color, forehead color, head pattern
forehead	forehead color, head pattern
left/right eye	eye color, head pattern
left/right leg	leg color
left/right wing	wing color, wing shape, wing pattern
nape	nape color
tail	tail shape, upper tail color, under tail color, tail pattern
throat	throat color, head pattern

Table 2: Attributes assignments on CUB200 [7]

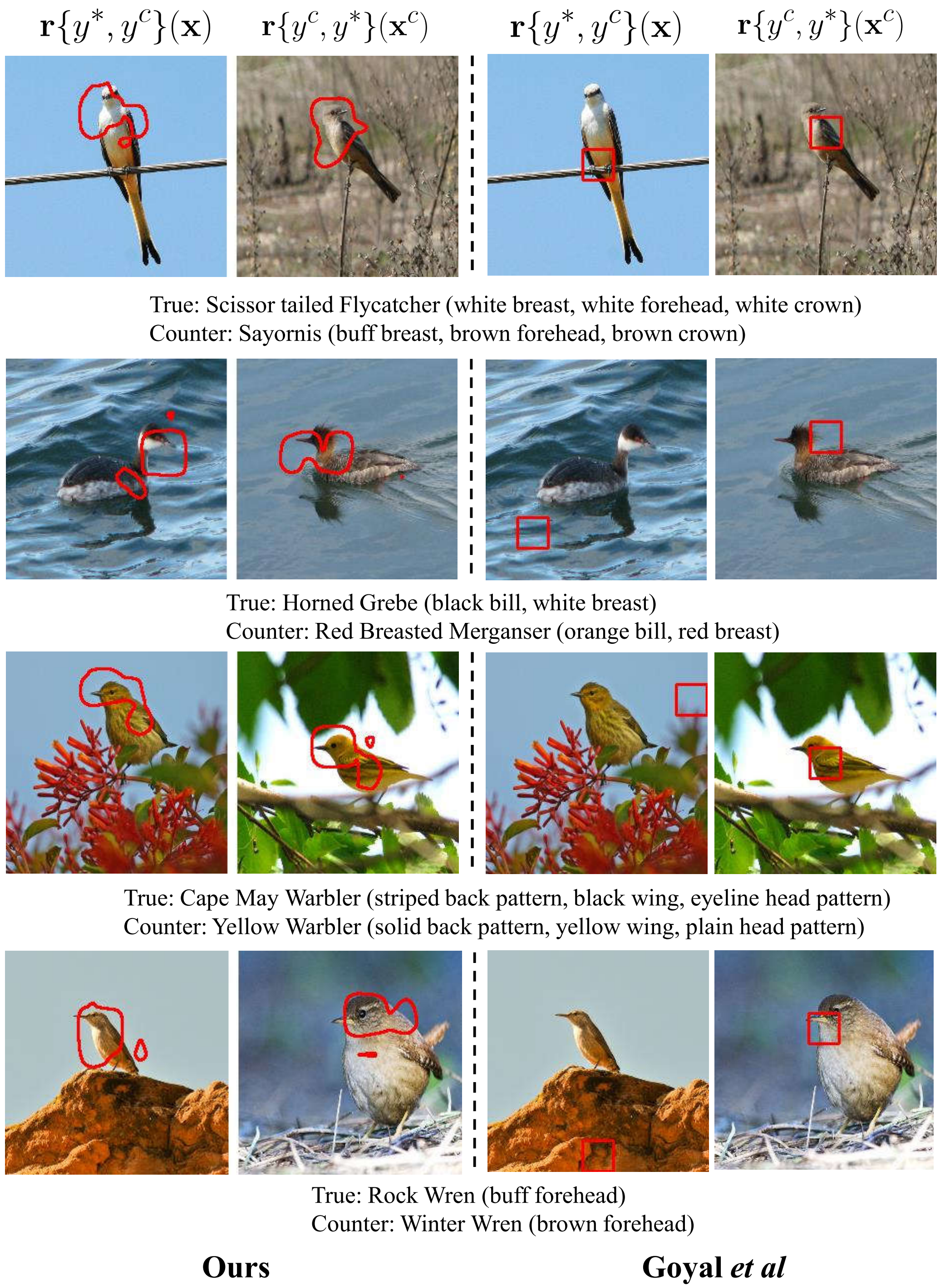

Figure 1: Comparison of counterfactual explanations (true and counter classes shown below each example, and ground truth class-specific part attributes in parenthesis).

Figure 2: Counterfactual explanations on CUB200 (true and counter classes shown below each example, and ground truth class-specific part attributes in parenthesis).

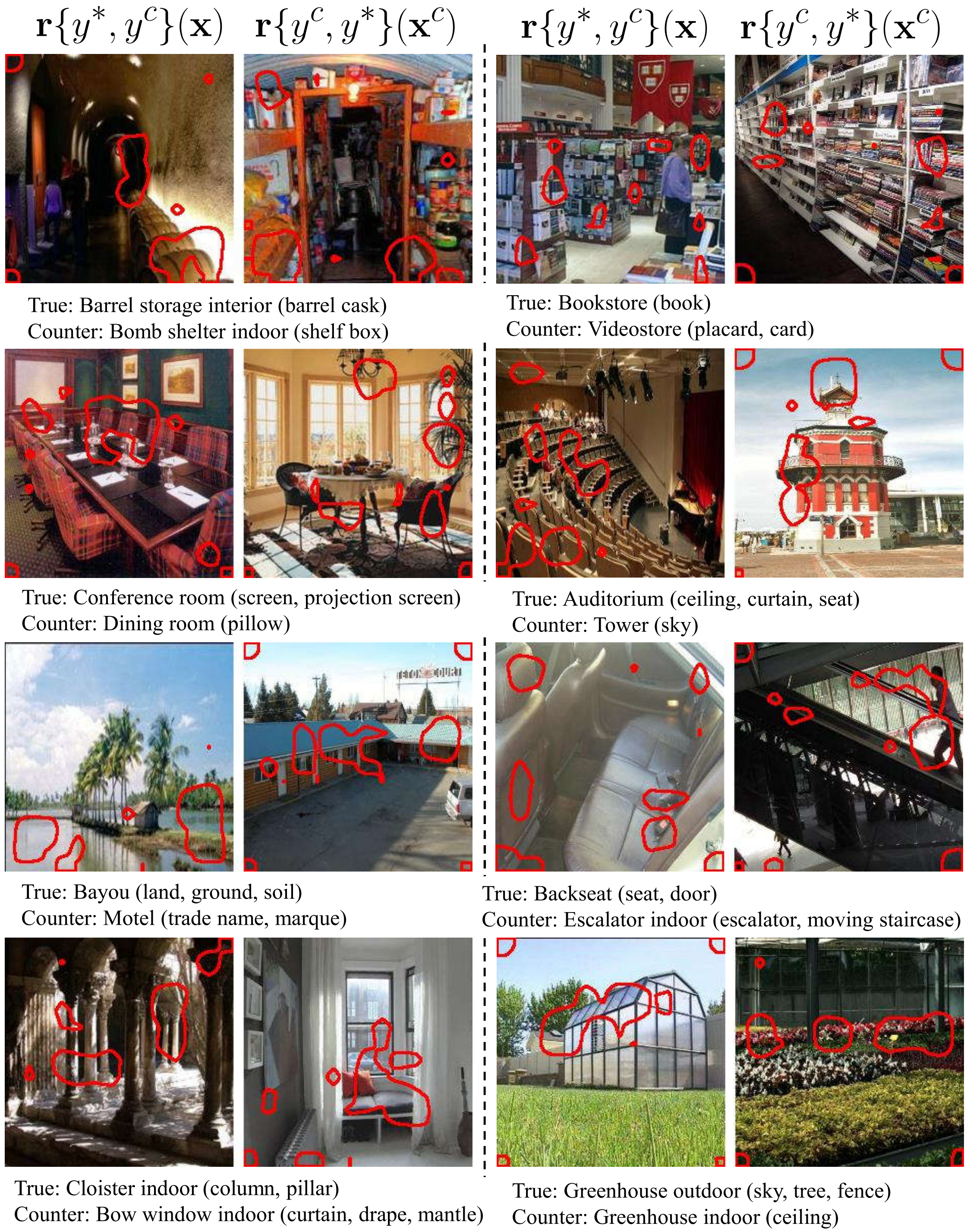

Figure 3: Counterfactual explanations on ADE20K.