A. Training details

A.1. Hyperparameters

Here we describe the hyperparameters to train our model in Table 8. We follow the training details in the original paper accompanied with these datasets as close as possible which is why there are some differences of hyperparameters among datasets. We set our maximum epochs to be large enough such that the performance saturates. We use floating point 16 to speed up the training. To our surprise in waterbirds when doing finetuning, floating point 16 is crucial to get superior performance and we use it for all our experiments. We release our code in https://github.com/zzzace2000/robust_cls_model.

Table 8: Training hyperparameters for each dataset

	IN-9	Waterbirds	Caltech Camera Trap
Model	BiT-S-R50x1	BiT-S-R50x1	BiT-S-R50x1
Epochs	25	40	50
LR	0.05	8.00E-04	0.003
Optimizer	SGD with momentum 0.9	SGD with momentum 0.9	RMSProp with momentum 0.9
Weight decay	1e-4	1.00E-04	0
LR Scheduling	Decay 1/10 after 6, 12, 18 epochs	Reduce LR on Plateau with patience 1	Decay 1/10 after 15, 30, 45 epochs
Batch size	32	32	64
Early stopping	Yes	Yes	Yes
Finetuning	No	Yes	No

A.2. Tiled background generation

Here we show how to generate tiled background in Algorithm 1.

Algorithm 1: Tiled background generation

Input: An image x and important region r
Output: Tiled image ϕ_{tile}
$A \leftarrow$ the largest rectangular regions that $r = 0$
$w, h \leftarrow x$ width, x height
$a_w, a_h \leftarrow A$ width, A height
if $a_w < w$ **then**
  repeat A $\lceil \frac{w}{a_w} \rceil$ times horizontally
end if
if $a_h < h$ **then**
  repeat A $\lceil \frac{h}{a_h} \rceil$ times vertically
end if
$A \leftarrow A[: w, : h]$
$\phi_{tile} = x\dot{r} + \dot{A}(1 - r)$

B. Additional results

B.1. Qualitative examples

We *randomly* pick images in IN9 Mixed-Next test set that our best model (CF(CAGAN)+F(Shuffle)+Sal) predicts correctly while Original model fails in Figure 10. We also *randomly* pick images in CCT Trans test set that our best model (CF(Tile)+Sal) predicts correctly while Original model fails in Figure 11.

Figure 10: Random examples that the best model (Grey) (CF(CAGAN)+F(Shuffle)+Sal) predicts correctly while Original model (Red) fails in IN9 Mixed-Next dataset. Here we show their top 5 predictions.

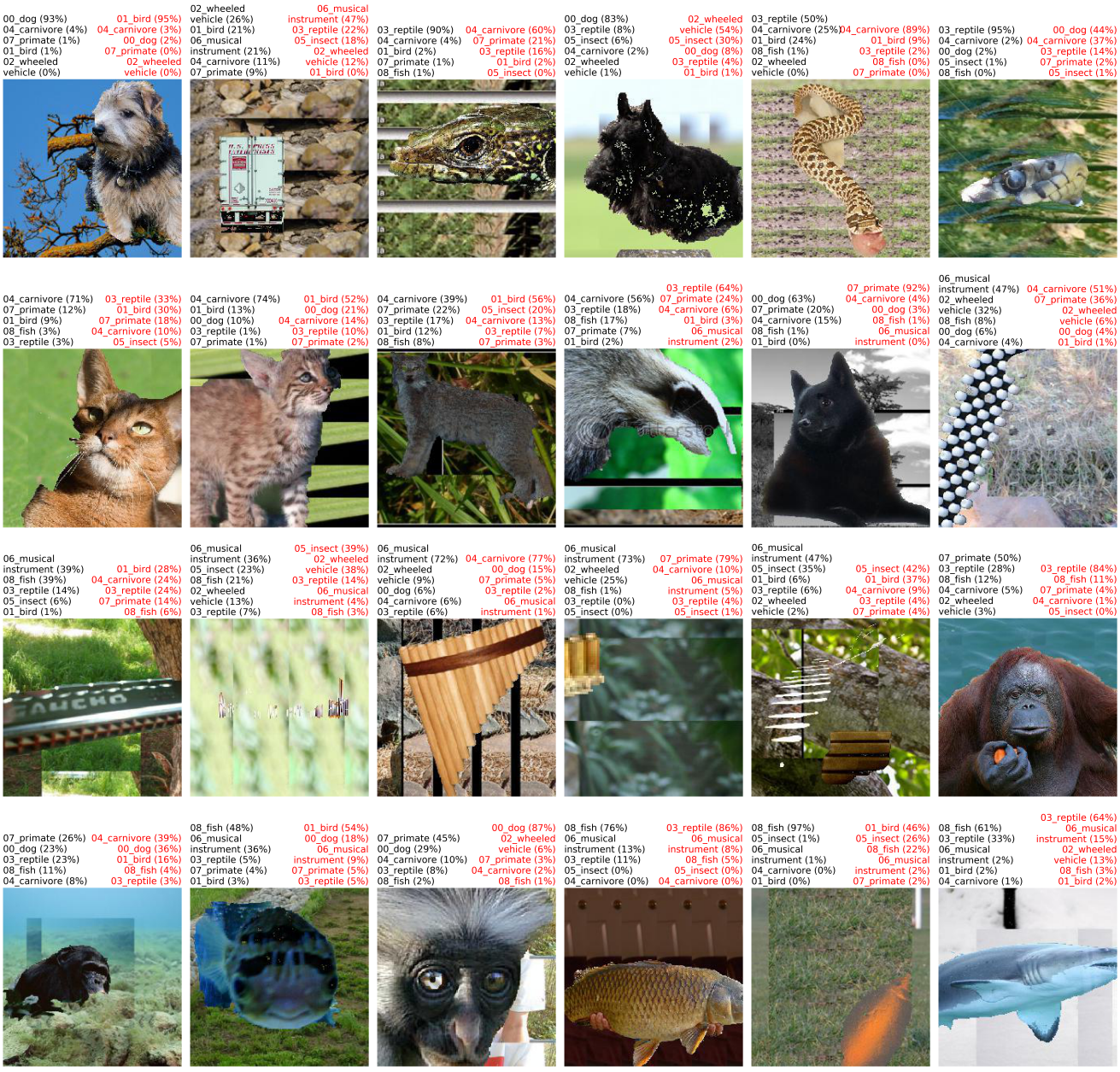

Figure 11: Random examples of CCT Trans-Test that the best model (+CF(Tile)+Sal) predicts correctly but Original model fails. The grey and red text are the top 5 predictions from the best and Original model's predictions respectively.