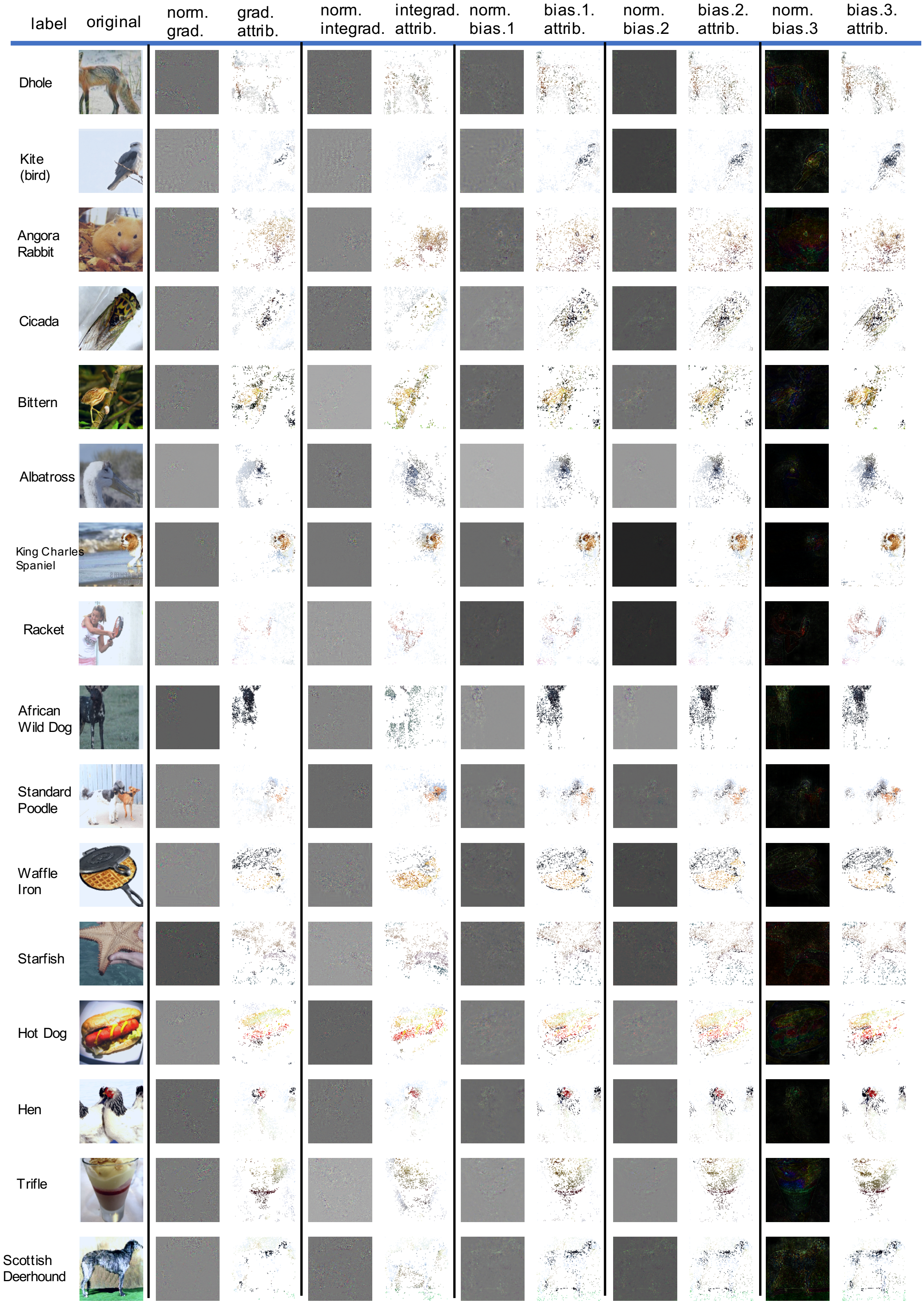

Figure 5: More visualizations of bias attribution on the ImageNet compared to gradient and integrated gradient attribution.

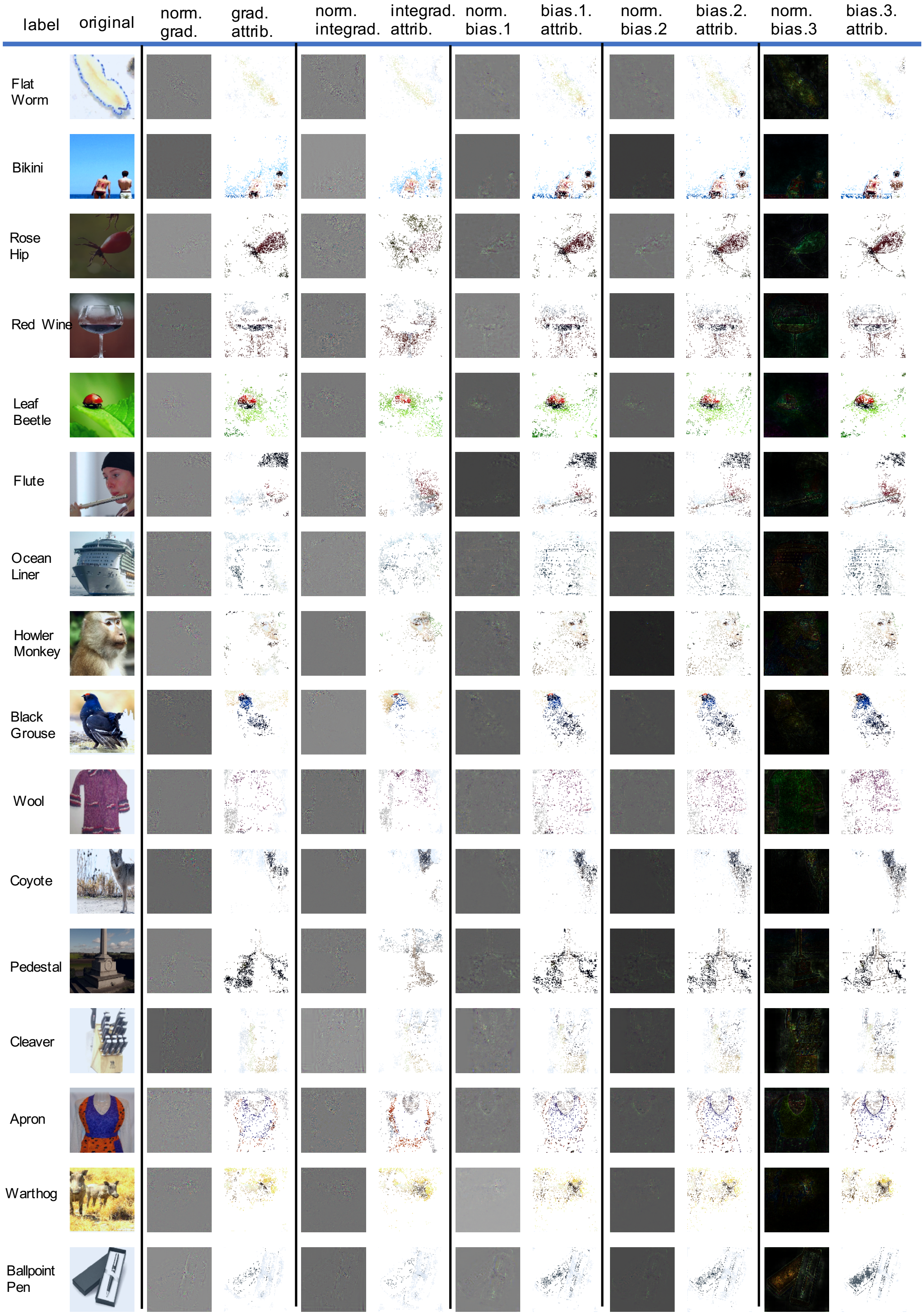

Figure 6: More visualizations of bias attribution on the ImageNet compared to gradient and integrated gradient attribution.

A MNIST Digit Flip Test Network Details

The network we use is a slight modification of the lenet structure (see Table 2). The trained network achieves 91.1% accuracy on the test set. For the digit flip test, we remove 120 pixels from each image to convert from the source class to the target class based on the descending order of source class attribution scores minus target class attribution scores. For the bias attribution methods, since the attribution scores add up to the bias term, which is different for different classes, we normalize the bias attribution scores by their L2 norm.

Layer Type	Layer Size (kernel size, stride, padding, channels) / # hidden units
conv1	[3 × 3], 1, 1, 32
Batchnorm	N/A
ReLU	N/A
conv2	[3 × 3], 1, 1, 32
Batchnorm	N/A
ReLU	N/A
MaxPool	[2 × 2], 2, 0, N/A
Fully Connected	128
Batchnorm	N/A
ReLU	N/A
Fully Connected	10

Table 2: Neural network