Supplementary Material

1 Details of CNN Structures

The AlexNet [14] based image CNN we use in the experiments is structured as follows:

```
conv 11 x 11 x 96, stride 2, valid padding
maxpool size 4, stride 2
```

```
  conv 5 x 5 x 192
  maxpool size 4, stride 2
```

```
  conv 3 x 3 x 384
  conv 3 x 3 x 256
  conv 3 x 3 x 256
  maxpool size 4, stride 2
```

```
fc 4000
fc 4000
sigmoid 10
```

Table 1: Image CNN, standalone

The image CNN takes image input of size $227 \times 227$. During training, a raw image is cropped to this size at random positions; during inference, we crop the center patch of images to match the size.

The coefficient CNN can have replaceable convolutional layers. When trained or evaluated separately, the full structure is as follows:

```
(F-B conv layers)
fc 256
fc 256
sigmoid 10
```

Table 2: Coefficient CNN, standalone

The joint network is structured as follows:

```
(Image conv layers) | (F-B conv layers)
concat
fc 4000
fc 4000
sigmoid 10
```

Table 3: Joint network

2 Experiments of F-B CNN Architectures

We devise several different network structures to run on Fourier-Bessel coefficients for comparison (listed in Table 4). Because the Fourier-Bessel images are not big (only $40 \times 11 \times 2$), we only investigate up to four convolutional layers. Our principle of probing coefficient network structures is to imitate the design of VGGNet [22], which is also the methodology
of He et al.’s experiments of residual networks [7], i.e., we apply homogeneous $3 \times 3$ convolutions solely and double the number of output maps right after any max-pooling. Each network structure is applied to both raw F-B coefficients and log F-B coefficients via (10).

We run training and testing with these convolutional configurations, and record the mean average precisions (mAPs) of the 10 attributes in Table 5. Experimental results confirm the common belief that more convolutional layers lead to better performance, but specific processing such as max-pooling and pseudo-log is crucial for convergence.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>Log</td>
</tr>
<tr>
<td>1conv</td>
<td>0.7031</td>
</tr>
<tr>
<td>2conv</td>
<td>0.7305</td>
</tr>
<tr>
<td>3conv</td>
<td>0.6703</td>
</tr>
<tr>
<td>4conv</td>
<td>0.6649</td>
</tr>
<tr>
<td>2c-pool-2c</td>
<td>0.6817</td>
</tr>
</tbody>
</table>

Table 5: Mean Average Precision with Different Coefficient CNN Architectures. Raw: Fourier-Bessel coefficients are fed into coefficient CNN with no processing. Log: Pseudo-log (10) is performed.