Goal: Analyze the variation of features generated by CNNs with respect to scene factors that occur in images

Step 1: Generate controlled image set
Step 2: Extract CNN layer responses
Step 3: Analyze and visualize responses

Pre-trained CNNs
- AlexNet, Places, VGG-S, GoogLeNet

Stimuli Images
- 2D synthetic
  - Single-color images
  - Black rectangle on white background
  - Colored square on a constant colored background
  - 3D CAD rendered views (Princeton ModelNet)
  - ~2900 models of chairs, cars, sofas, toilets, beds

3D Scene Factors
- Style, Color, Orientation, Lighting, Position, Scale

Scene Factor Analysis
\[ F^L(\theta) = \sum_{k=1}^{N} F_k^L(\theta_k) + \Delta^L(\theta) \]

Relative variance of two scene factors: 2D position and aspect ratio of a black rectangle on white background

<table>
<thead>
<tr>
<th>Scene Factor</th>
<th>Relative Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D position</td>
<td>Aspect ratio</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
</tr>
<tr>
<td>AlexNet, pool5</td>
<td>49.8% 9.5% 40.8%</td>
</tr>
<tr>
<td>AlexNet, fc6</td>
<td>45.1% 22.3% 32.6%</td>
</tr>
<tr>
<td>AlexNet, fc7</td>
<td>33.9% 37.0% 29.1%</td>
</tr>
</tbody>
</table>

Other observations
- Relative to object style, color is more important for Places than for AlexNet and VGG. This difference is more pronounced for the background color than for foreground.
- VGG fc7 layer appears to be less sensitive to viewpoint than AlexNet and Places.
- In the last layers, the number of dimensions used for style is much larger than for viewpoint. This effect is more pronounced for AlexNet and VGG than for the Places.