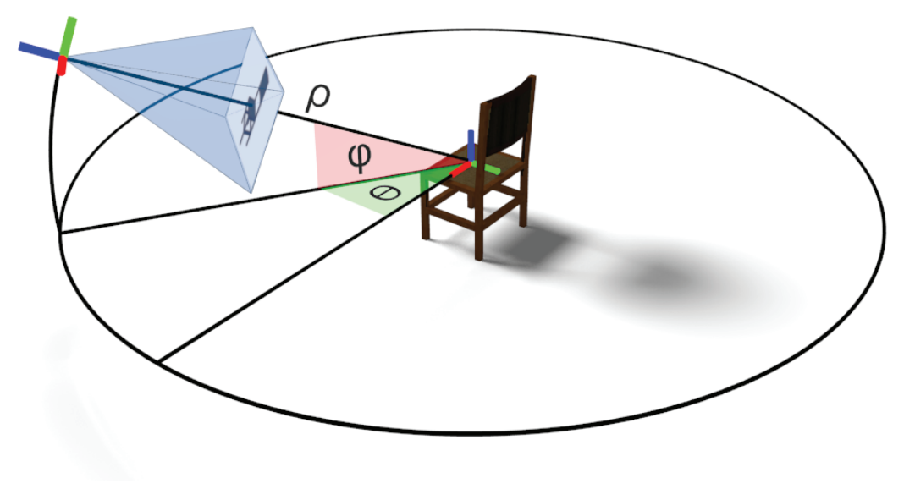
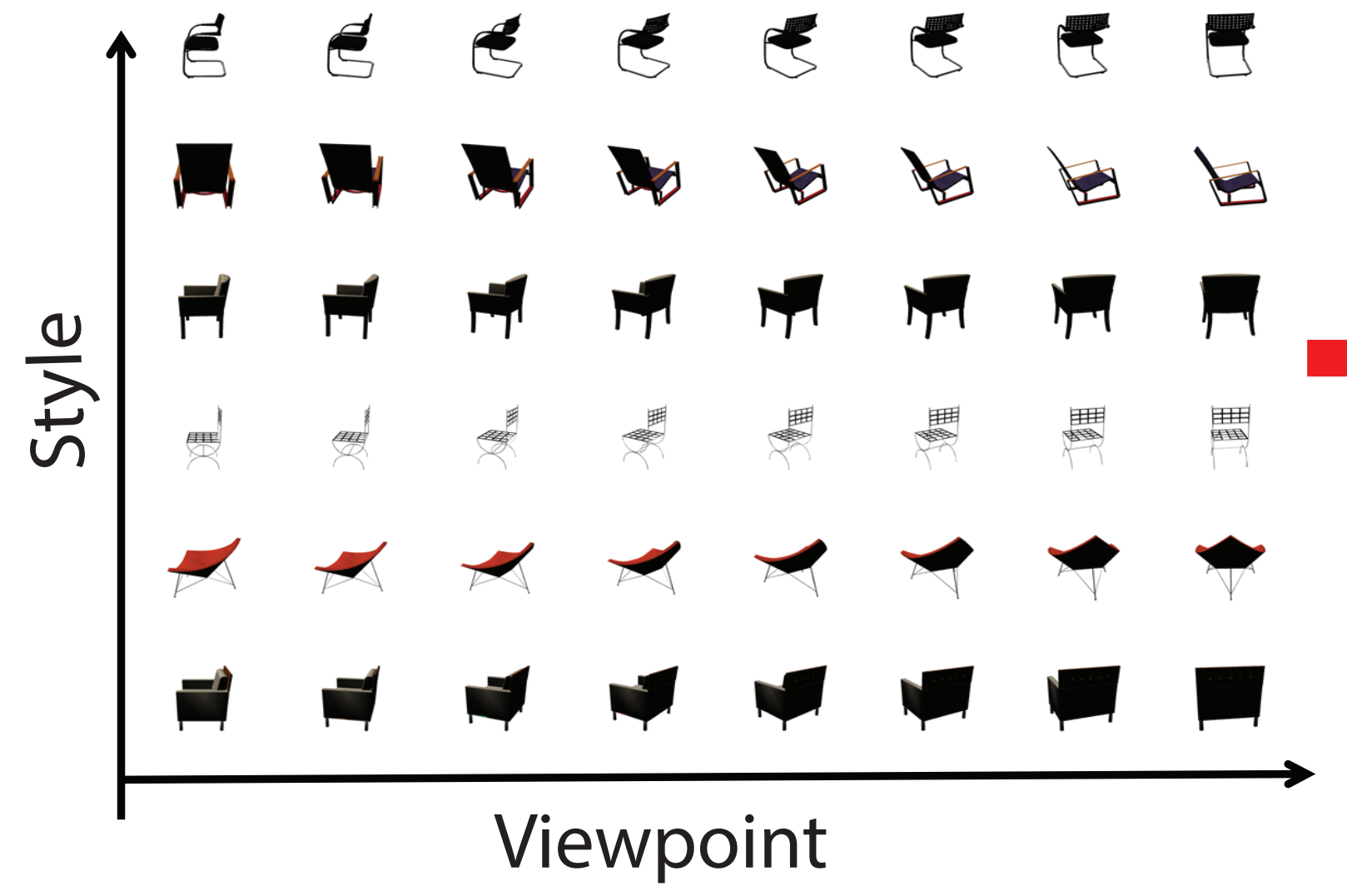
# Understanding Deep Features with Computer-Generated Imagery

Mathieu Aubry
ENPC ParisTech & UC Berkeley

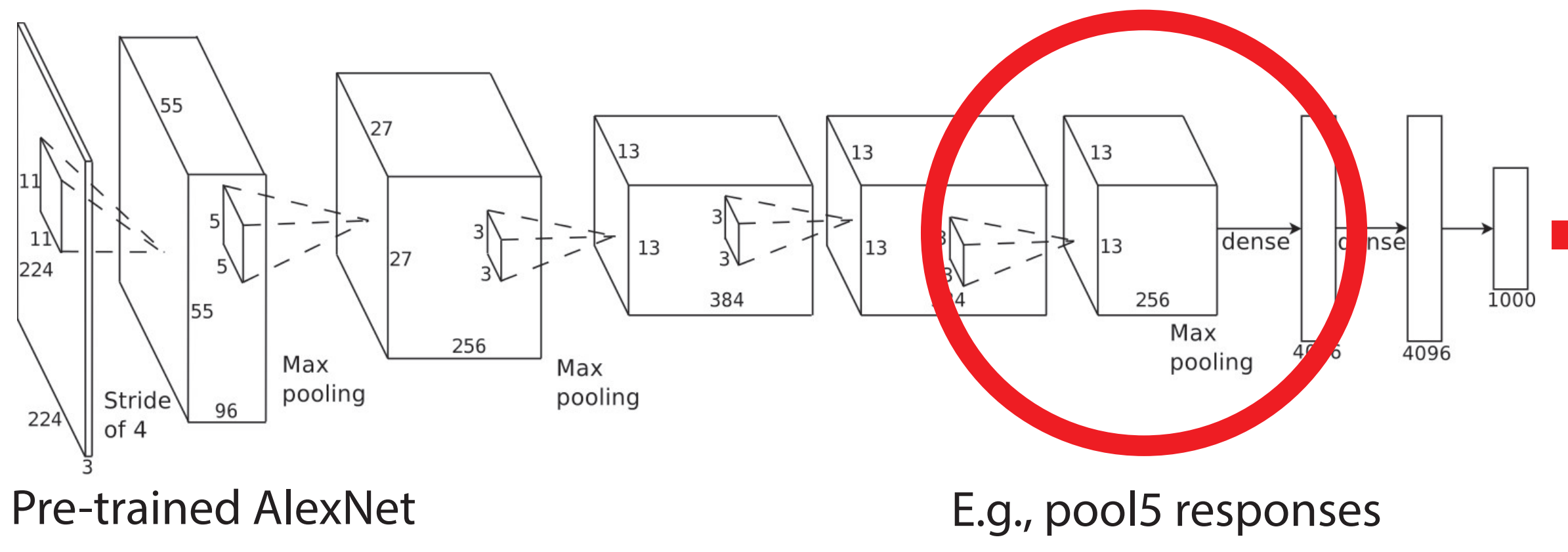
Bryan Russell
Adobe Research

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

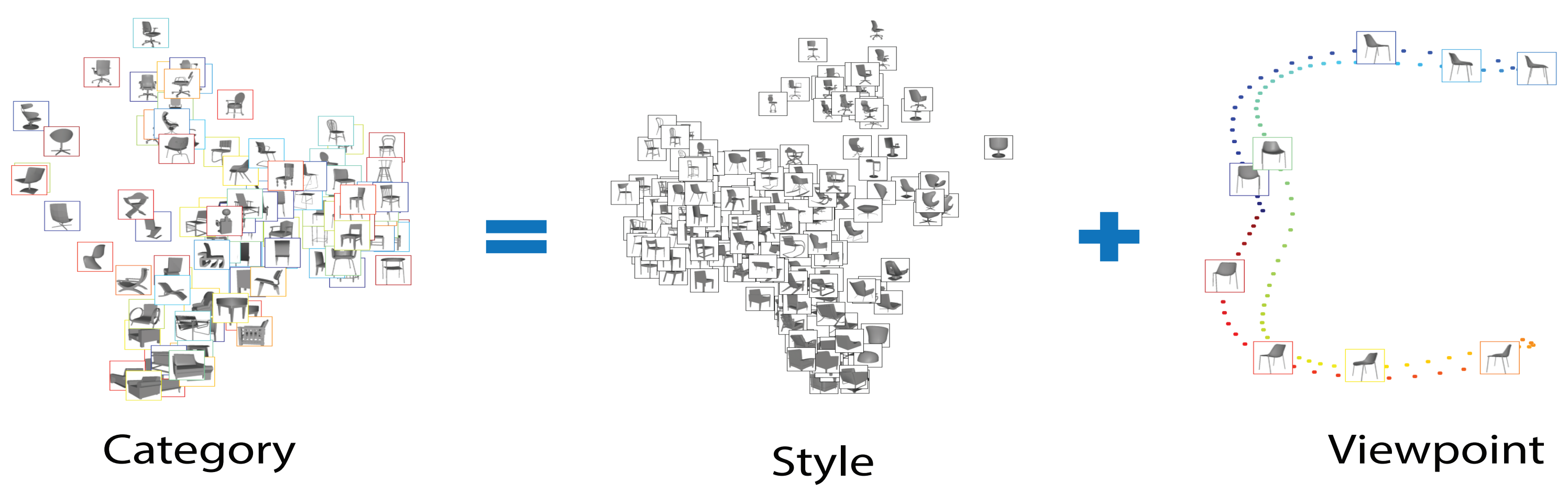




Step 1: Generate controlled image set



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

Rendering Parameters $\theta = (\theta_1, \ldots, \theta_N)$

E.g. (azimuth angle, CAD ID)

$$F^L(\theta)$$

Centred CNN Layer Responses (E.g. AlexNet "pool5")

$$F_k^L(t) = \mathbb{E}(F^L(\theta)|\theta_k = t)$$

Marginalized Feature for Factor k — Centred CNN Layer Responses

$$F^L(\theta) = \sum_{k=1}^{N} F_k^L(\theta_k) + \Delta^L(\theta)$$

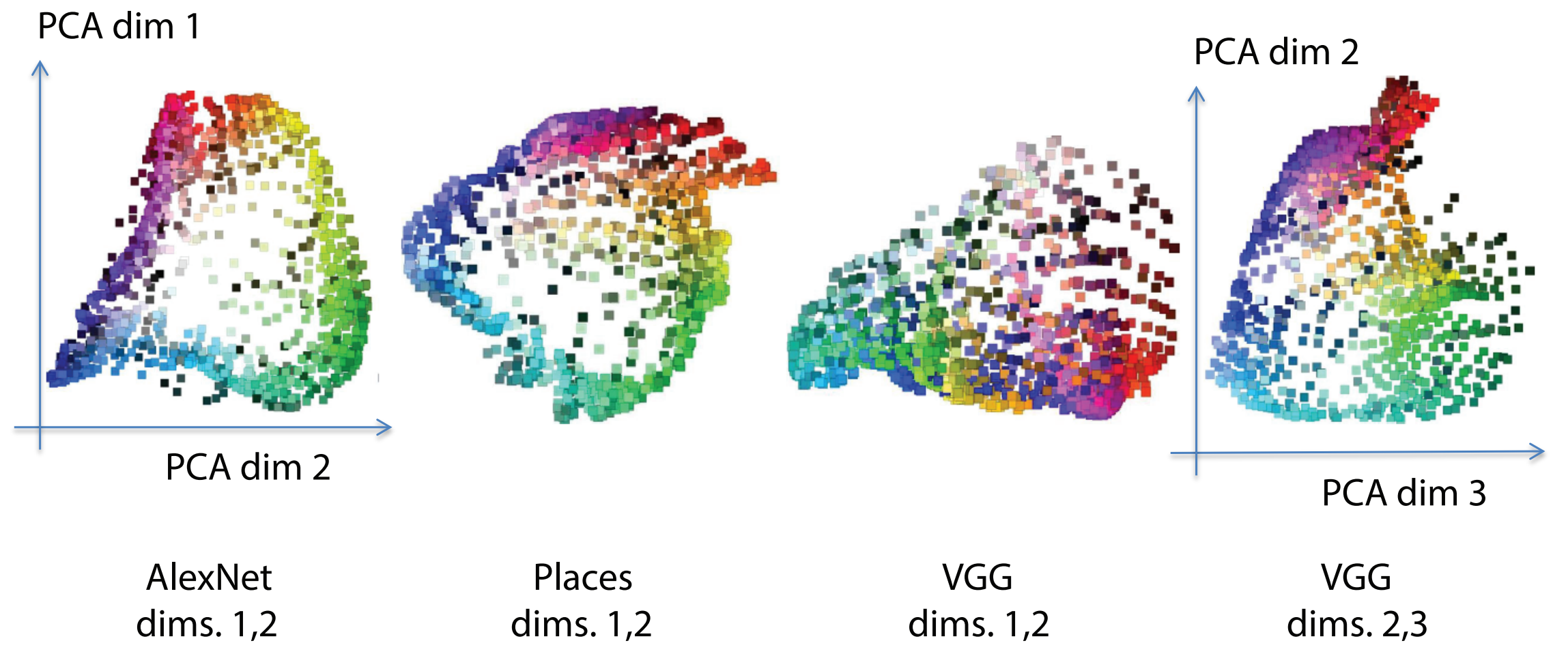
Centred CNN Layer Responses — Marginalized Features — Residual

$$\sum_{k=1}^{N} \frac{\mathrm{var}(F_k^L)}{\mathrm{var}(F^L)} + \frac{\mathrm{var}(\Delta^L)}{\mathrm{var}(F^L)} = 1$$

Scene Factor Relative Variances — Residual Relative Variance

## 2D stimuli

**PCA of AlexNet fc7 of one scene factor:** color of single-colored images

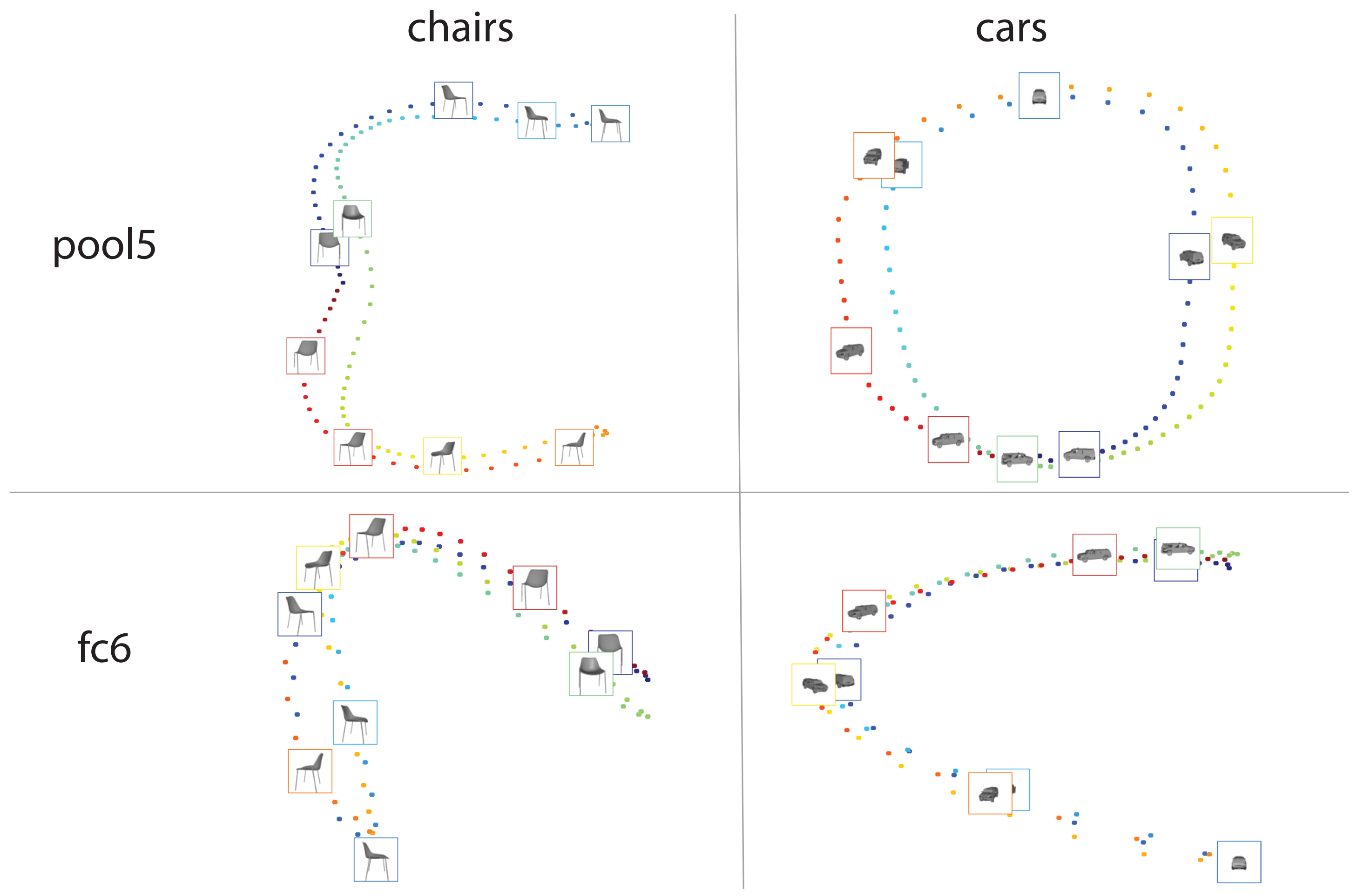


**Relative variance of two scene factors:** 2D position and aspect ratio of a black rectangle on white background

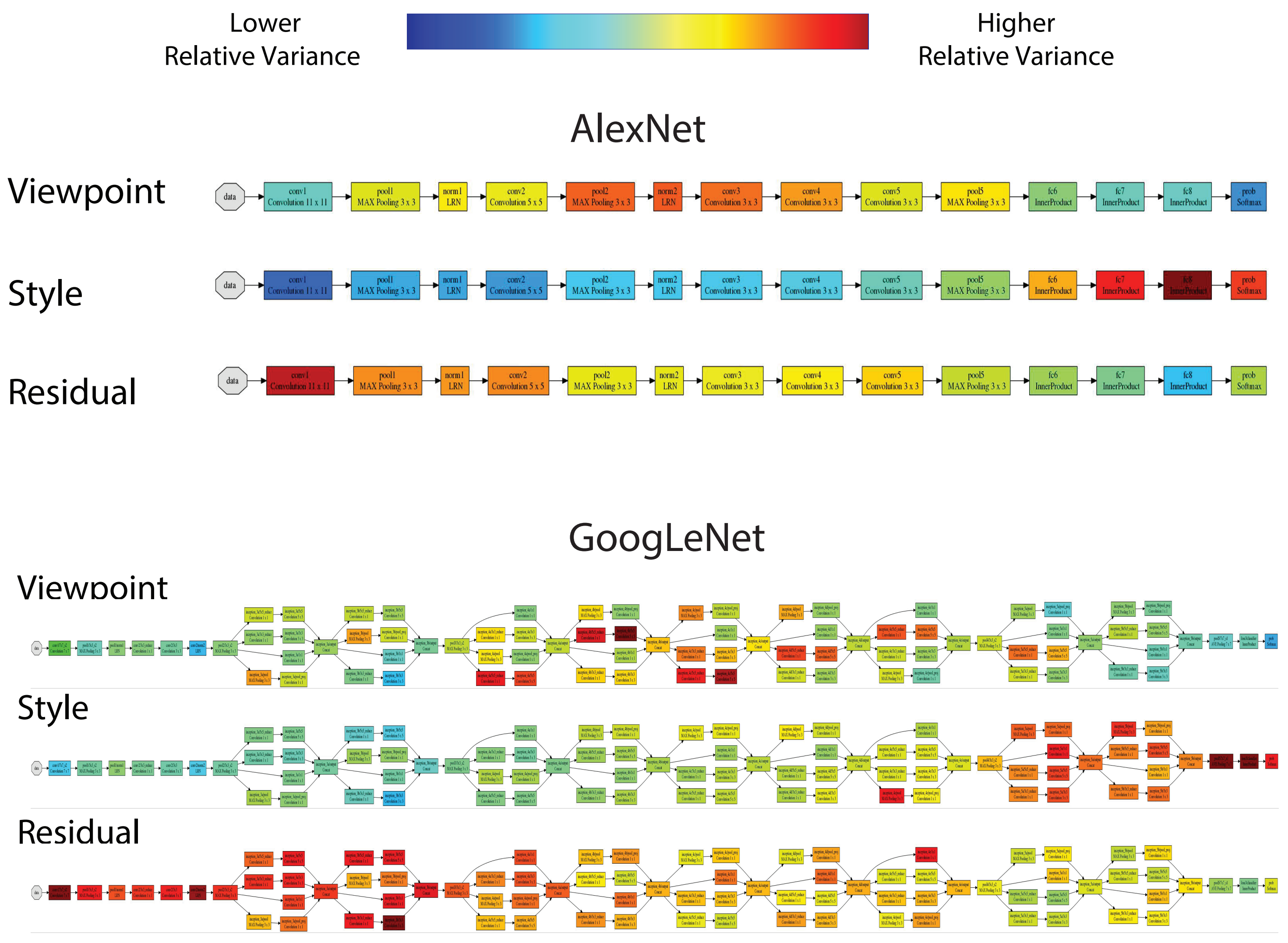
| | 2D position | Aspect ratio | Residual |
|---|---|---|---|
| AlexNet, pool5 | 49.8 % | 9.5 % | 40.8 % |
| AlexNet, fc6 | 45.1 % | 22.3 % | 32.6 % |
| AlexNet, fc7 | 33.9 % | 37.0 % | 29.1 % |

## Qualitative comparison of rotation embedding

AlexNet category rotation embedding (PCA)



## Full network quantitative analysis



## Relation to real photographs

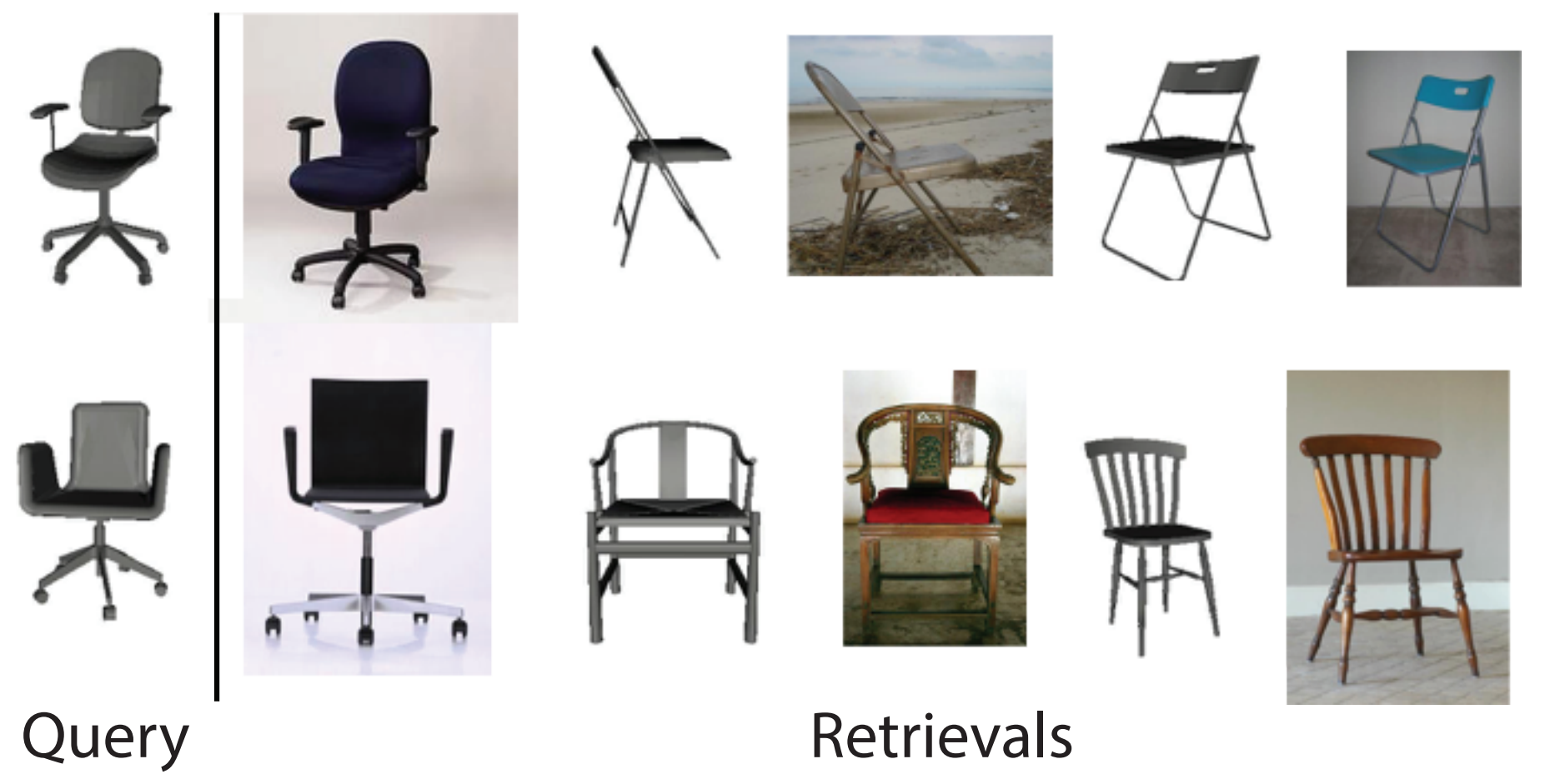
### Similar results on a smaller scale on ETH-80

- Dataset consists of toys or small objects
- 8 categories:
  apples, pears, tomatoes, cows, dogs, horses, cups, cars
- 10 instances per category, 41 viewpoints

| | Rotation | Style | Residual |
|---|---|---|---|
| AlexNet, pool5 | 35.4 % | 21.6 % | 43.0 % |
| AlexNet, fc6 | 30.2 % | 27.7 % | 42.0 % |
| AlexNet, fc7 | 29.5 % | 30.5 % | 40.0 % |

### Cross-domain nearest neighbors

Dot-product similarity over AlexNet pool5 features



## 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.