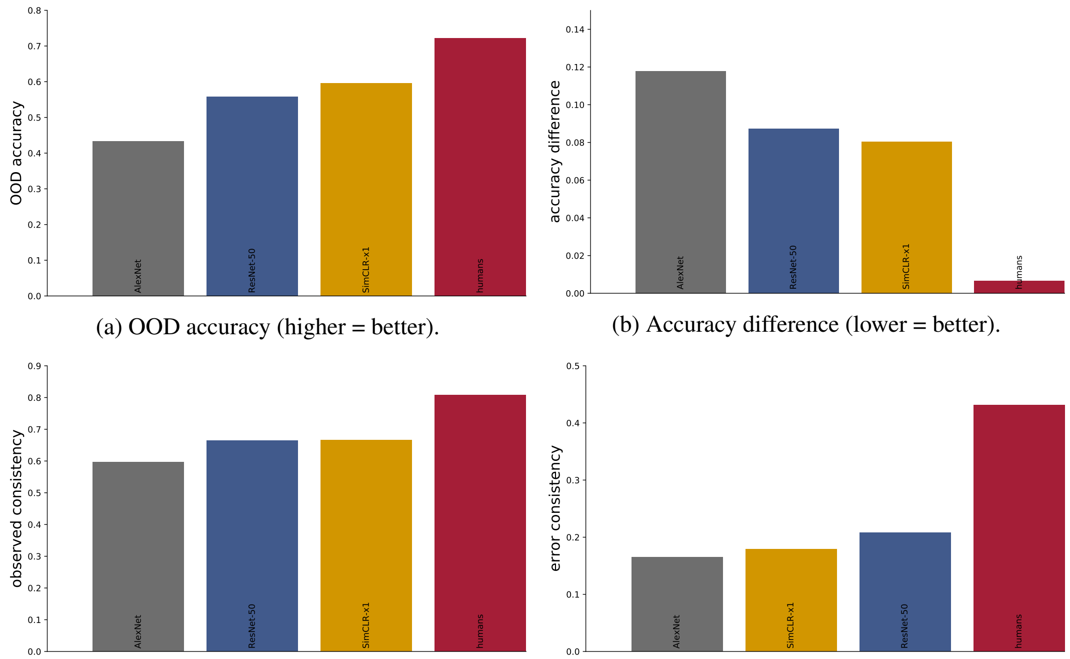
Generalisation report

Generalisation report produced by model-vs-human



(a) OOD accuracy (higher = better).

(b) Accuracy difference (lower = better).

(c) Observed consistency (higher = better).

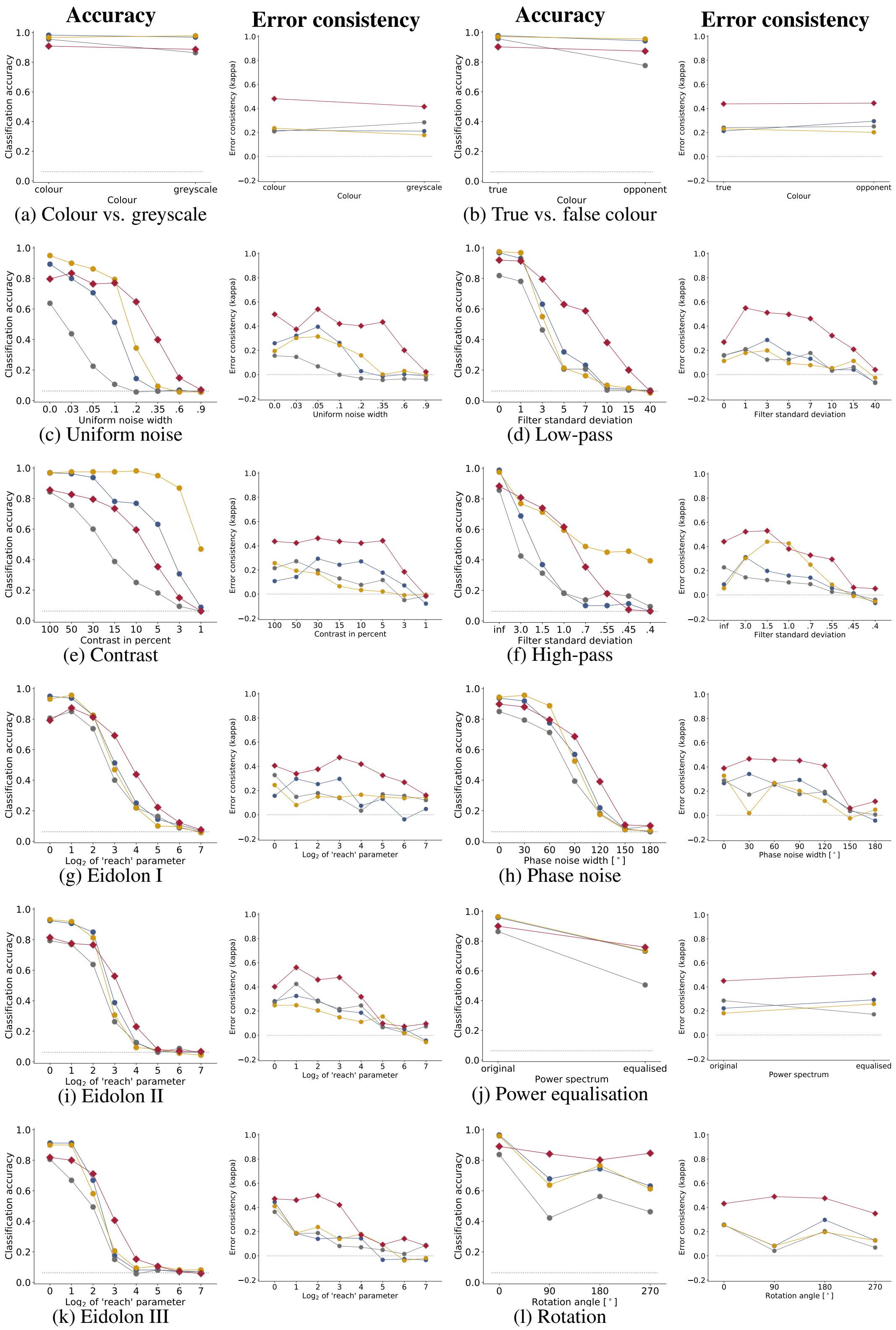
(d) Error consistency (higher = better).

Figure 1: Benchmark results for different models, aggregated over datasets.

Table 1: Benchmark table of model results for most human-like behaviour. The three metrics “accuracy difference” “observed consistency” and “error consistency” (plotted in Figure 1) each produce a different model ranking. The mean rank of a model across those three metrics is used to rank the models on our benchmark.

model	accuracy diff. ↓	obs. consistency ↑	error consistency ↑	mean rank ↓
SimCLR-x1	**0.080**	**0.667**	0.179	**1.333**
ResNet-50	0.087	0.665	**0.208**	1.667
AlexNet	0.118	0.597	0.165	3.000

Generalisation report.



(a) Colour vs. greyscale

(b) True vs. false colour

(c) Uniform noise

(d) Low-pass

(e) Contrast

(f) High-pass

(g) Eidolon I

(h) Phase noise

(i) Eidolon II

(j) Power equalisation

(k) Eidolon III

(l) Rotation

Figure 2: OOD accuracy and error consistency.

Table 2: Benchmark table of model results for highest out-of-distribution robustness.

model	OOD accuracy ↑	rank ↓
SimCLR-x1	**0.596**	**1.000**
ResNet-50	0.559	2.000
AlexNet	0.434	3.000

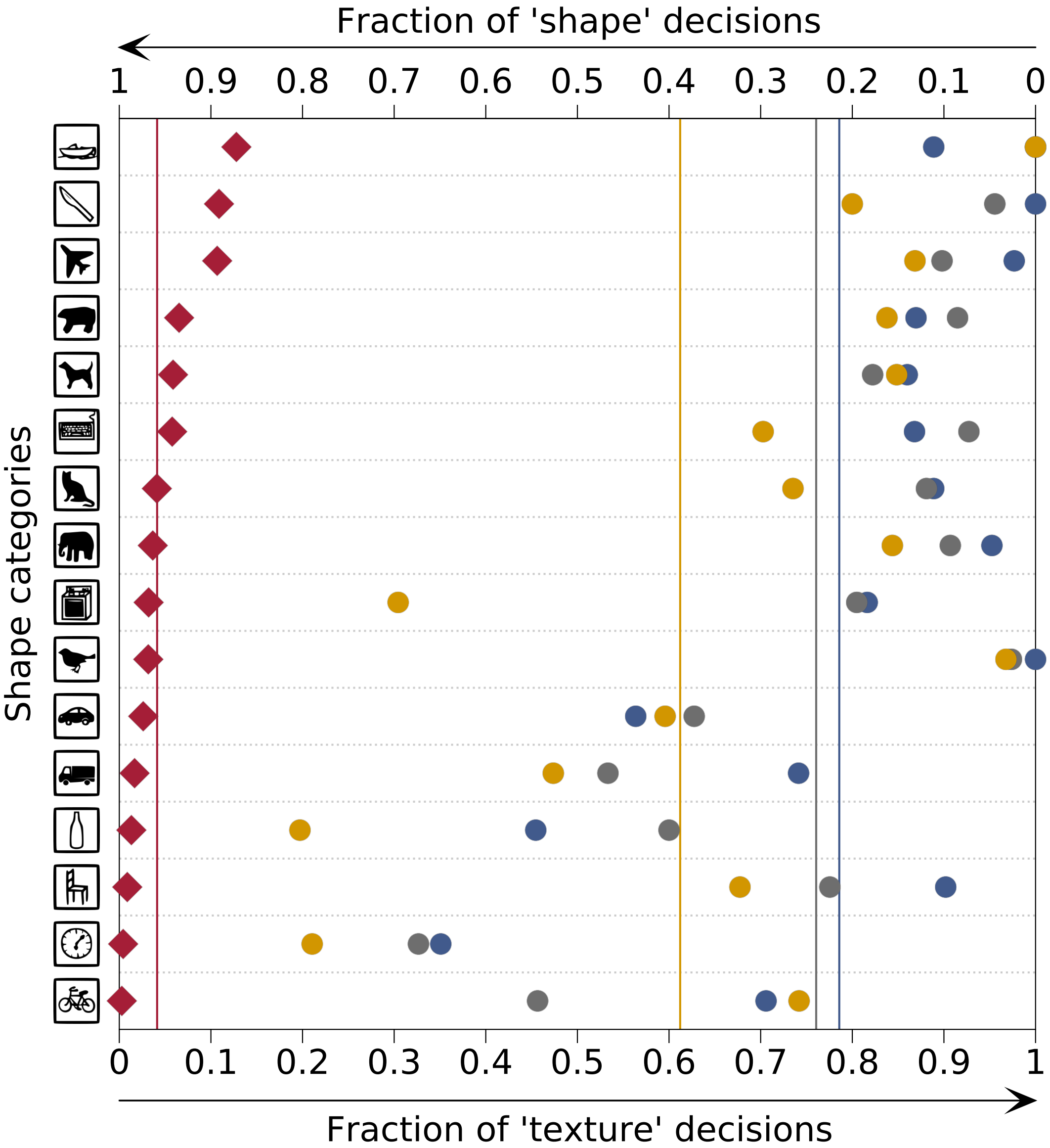


Figure 3: Shape vs. texture bias: category-level plot.

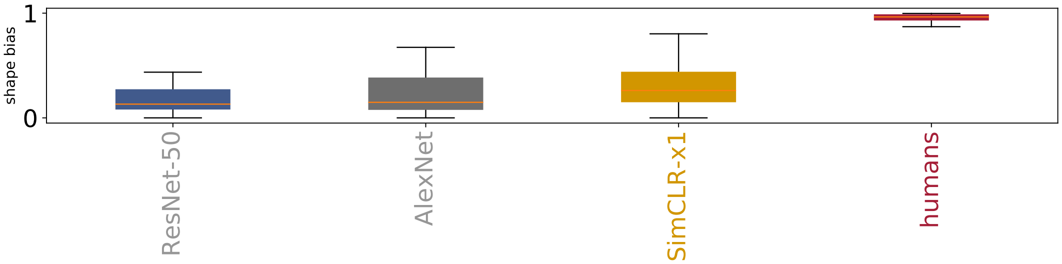
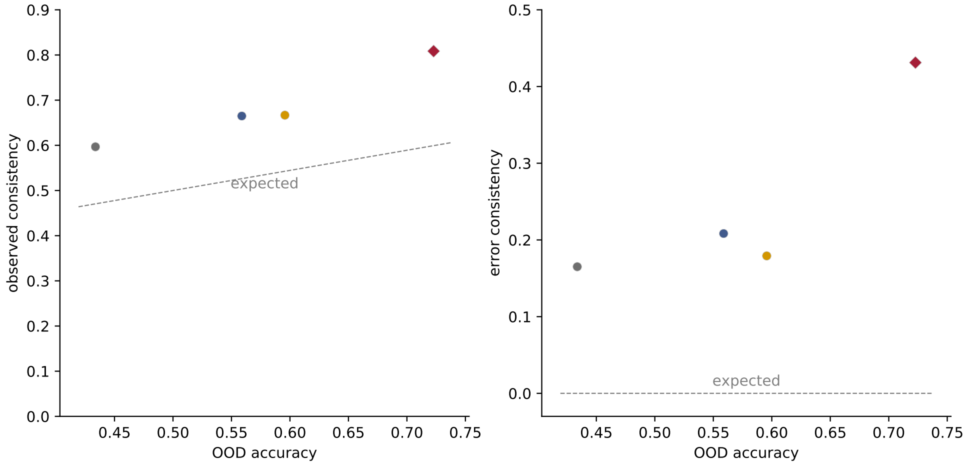


Figure 4: Shape vs. texture bias: boxplot.



(a) Out-of-distribution accuracy vs. observed consistency

(b) Out-of-distribution accuracy vs. error consistency

Figure 5: Observed consistency and error consistency between models and humans as a function of out-of-distribution (OOD) accuracy. Dotted lines indicate consistency expected by chance.

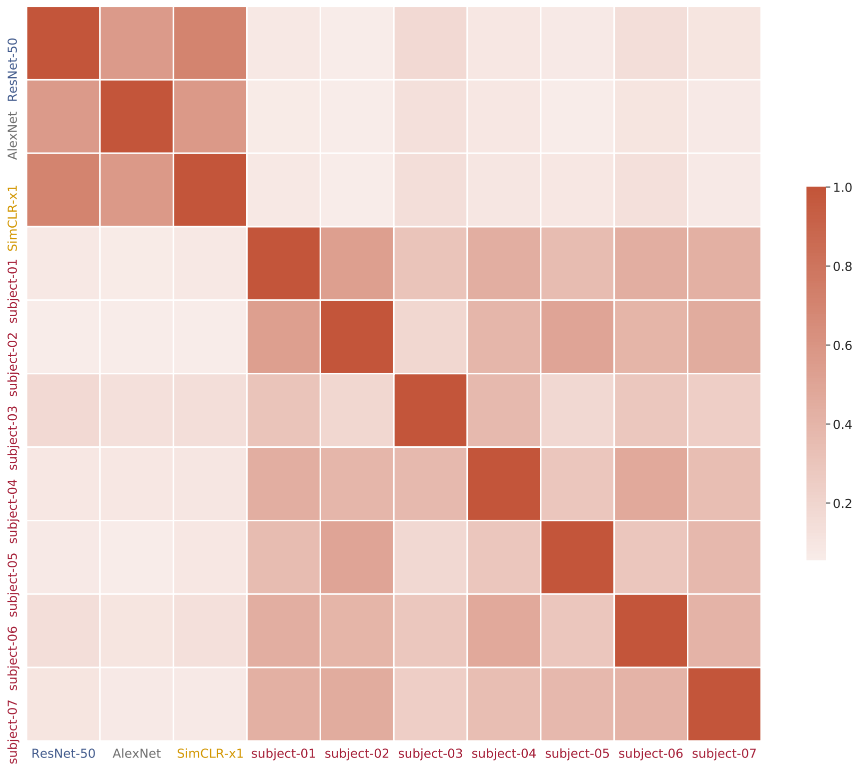


Figure 6: Error consistency for ‘sketch’ images.

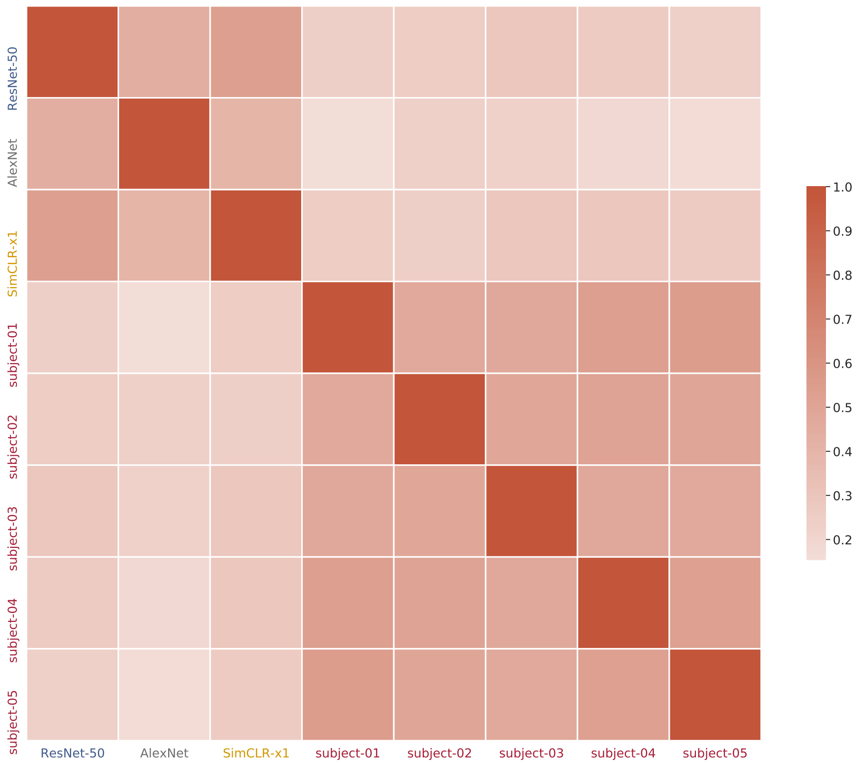


Figure 7: Error consistency for ‘stylized’ images.

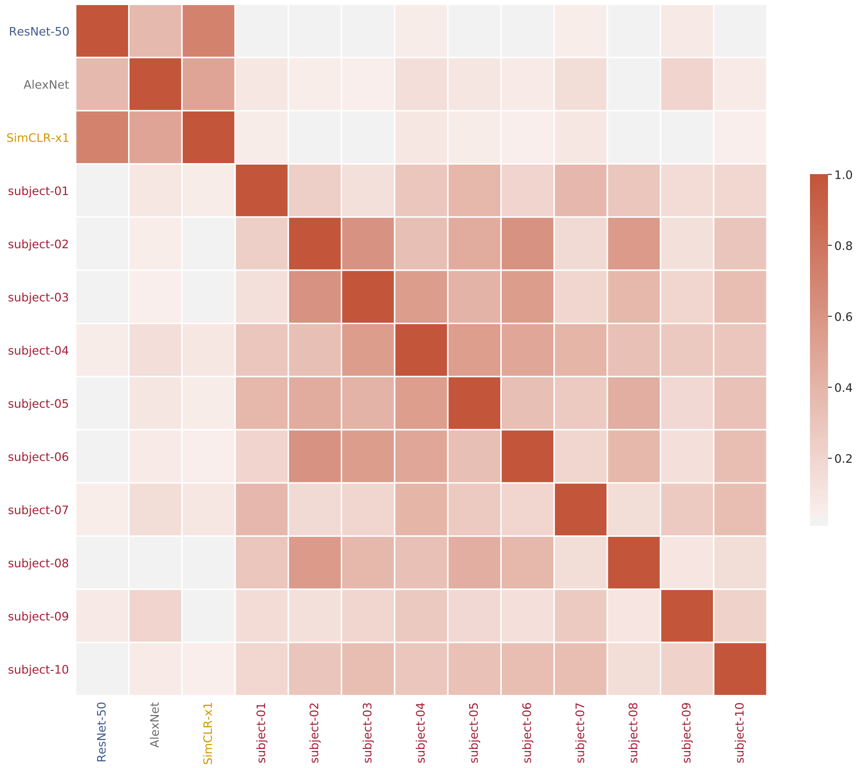


Figure 8: Error consistency for ‘edge’ images.

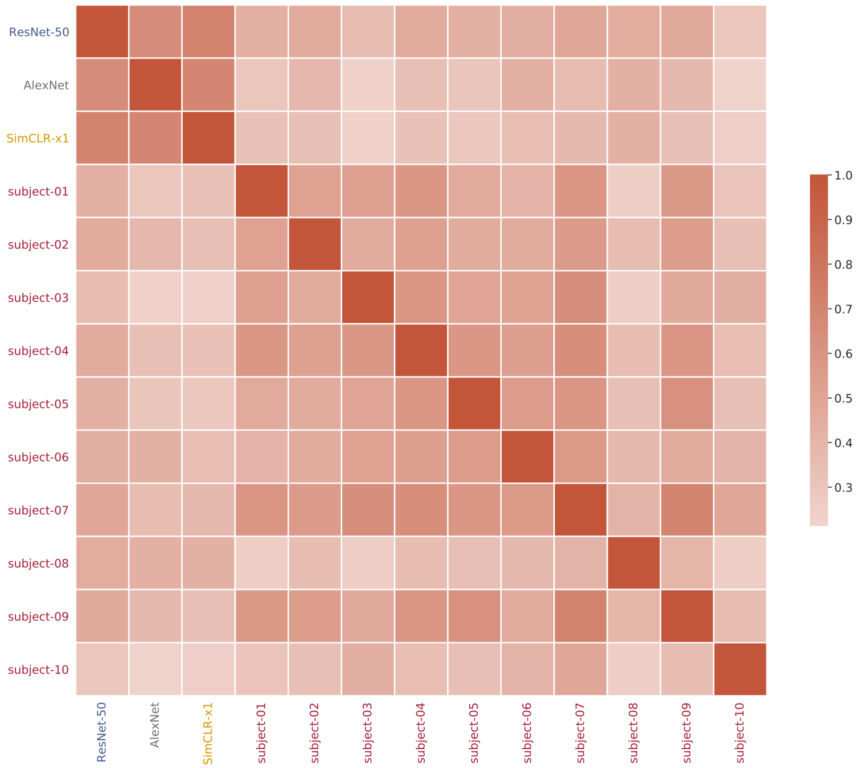


Figure 9: Error consistency for ‘silhouette’ images.

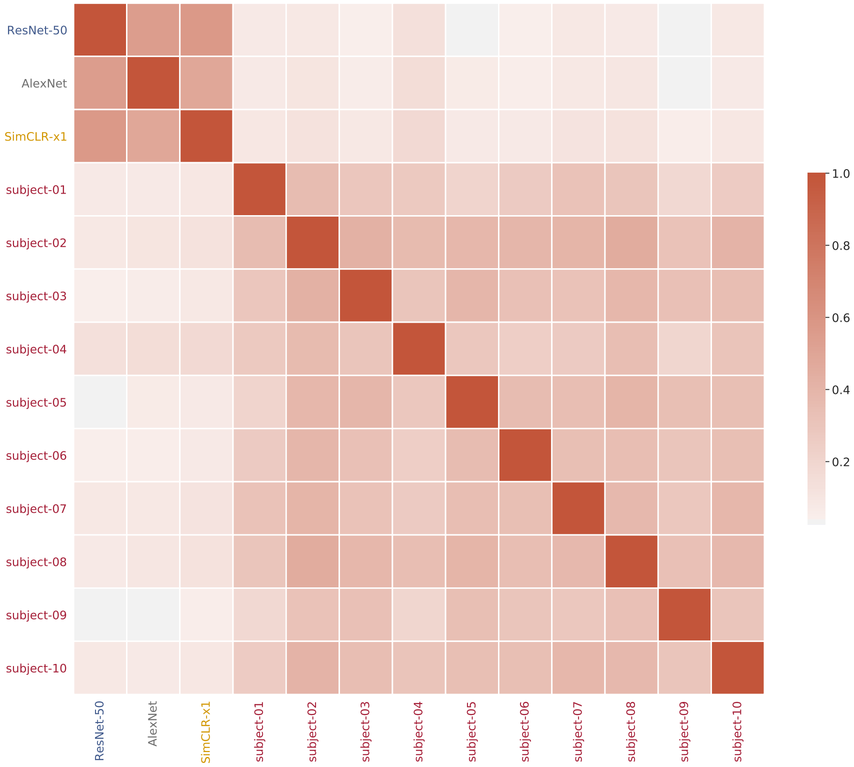


Figure 10: Error consistency for ‘cue conflict’ images.