Technical Appendix for 'Correct for Whom? Subjectivity and the Evaluation of Personalized Image Aesthetics Assessment Models'

Samuel Goree,[1] Weslie Khoo, [2] David Crandall [1]

[1] Department of Informatics, Indiana University
[2] Department of Computer Science, Indiana University
sgoree@iu.edu, weskhoo@iu.edu, djcran@indiana.edu

Threshold Estimation

When evaluating the accuracy of single-image scores for pairwise image labels, we use a threshold to separate out score differences into pairwise labels. We chose the value of the threshold empirically after the fact to maximize both accuracy values. We would emphasize again, however, that the AADB scores and scores inferred from PR-AADB cannot be compared fairly on this metric, since the latter are derived from the labels we are using to test them. The claims in our paper do not rely on such a comparison.

Full Regression Results

In Tables 1,2 below, we report the full coefficients of our regression model, fitted using Python statsmodels. Since these are logistic regression coefficients, they should be interpreted as effects on the log-odds ratio, e.g. a coefficient of 5 for a binary variable would mean that the presence of that variable increases the log-odds of consistency by 5, which corresponds to a multiplicative increase by e^5 in the odds of consistency, conditioned on the features. The columns correspond to:

- Feature: the feature name
- N: The number of image pairs for which this feature is 1
- coef: the coefficient value.
- std err: the standard error for the coefficient estimate
- z: the z score for the coefficient estimate used to produce the P value
- $P > |z|$: the P value, i.e. the probability of getting a parameter value this far from zero due to random chance.
- 0.025: The lower end of the 95% confidence interval for the parameter.
- 0.975: The upper end of the 95% confidence interval for the parameter.

Please see our main paper for details on how the aesthetic attribute and content features were computed.

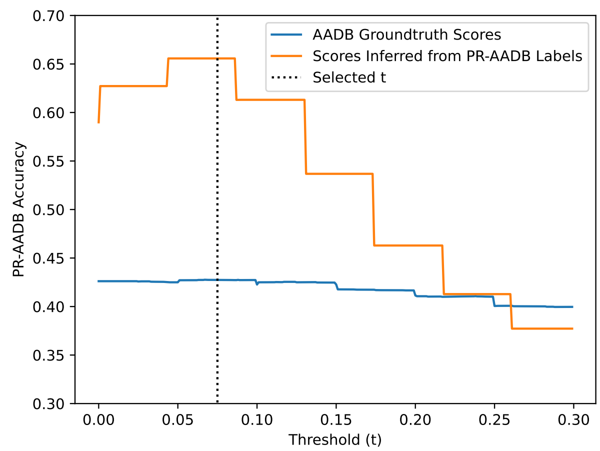

Figure 1: PR-AADB Accuracy Accross threshold values. We report results using the threshold 0.075, indicated with a dotted line, which maximizes accuracy.

Regression Results Split by Feature

We also conducted logistic regression analyses for each set of features separately. For these analyses, the prediction target remains agreement on the image pair level, but instead of fitting a single model, we fit three separate regression models to demographic, aesthetic and content features, respectively. Resulting coefficients are shown in Figure 2. These results are very similar to the results from the main paper.

Confusion Matrices

In Figure 4 in the main paper, we show accuracy scores for each participant under four sets of labels: labels inferred from the AADB scores, labels inferred from scores inferred from the PR-AADB labels, labels predicted by a generic model and labels predicted by a few-shot personalized model.

In this section, we report confusion matrices for these four prediction experiments. All experiments are on a 3-way classification problem: will the user prefer image a (-1), image b (1) or neither (0).

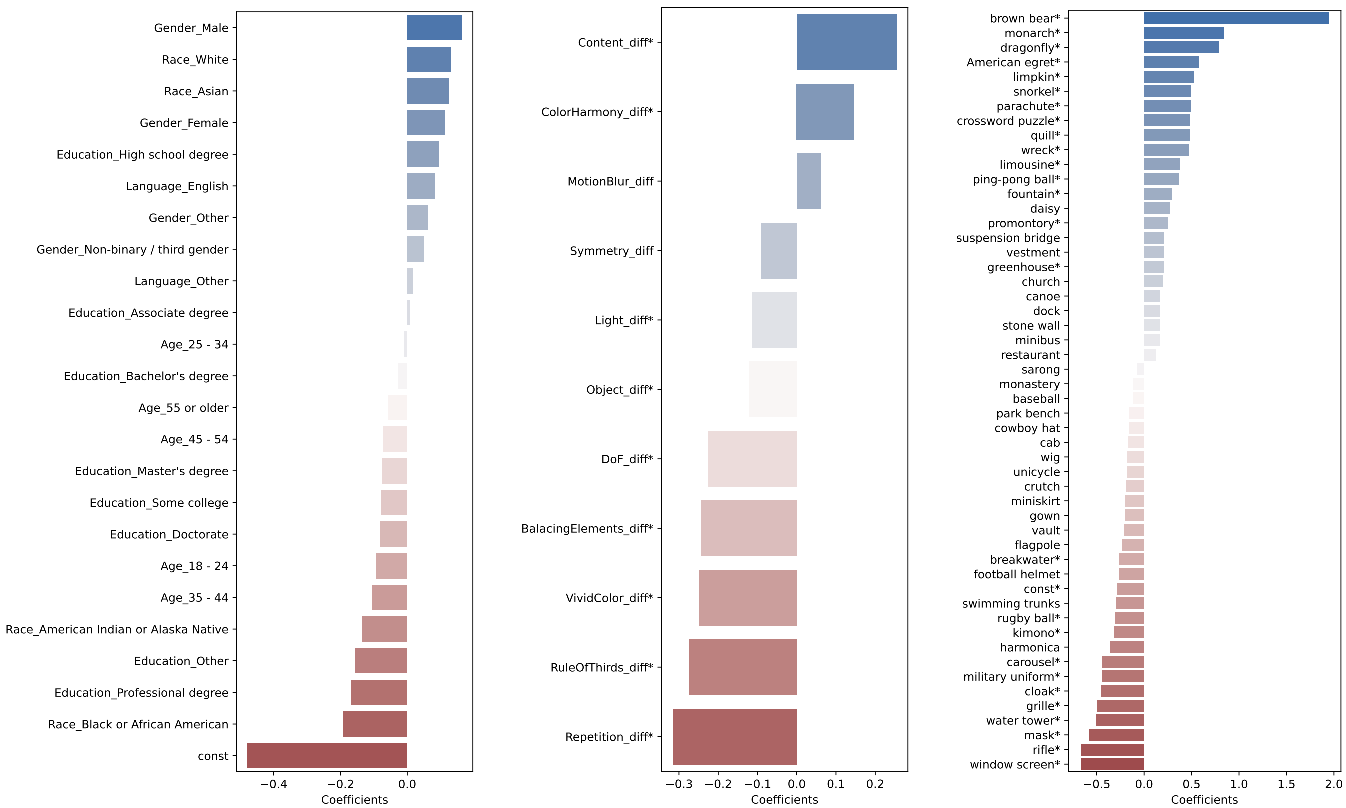

Figure 2: Coefficients for three separate regression analyses using each set of features. Stars on feature names indicate significance. Pseudo-R^2 values are 0.002, 0.007 and 0.015 respectively for the three analyses. Notably, none of the demographic coefficients are significantly different than zero.

Feature	N	coef	std err	z	P > |z|	0.025	0.975
const	12897	-0.8608	†	-6.67E-07	1	-‡	‡
Age 18 - 24	3948	-0.0272	0.351	-0.078	0.938	-0.715	0.661
Age 25 - 34	4777	0.0461	0.345	0.134	0.894	-0.63	0.722
Age 35 - 44	1085	-0.0537	0.349	-0.154	0.878	-0.738	0.63
Age 45 - 54	619	0.0166	0.357	0.047	0.963	-0.682	0.715
Age 55 or older	2390	0.0065	0.345	0.019	0.985	-0.67	0.683
Gender Female	6328	0.1152	0.242	0.477	0.634	-0.359	0.589
Gender Male	5369	0.1657	0.241	0.689	0.491	-0.306	0.637
Gender Non-binary/third gender	654	0.0247	0.259	0.095	0.924	-0.483	0.532
Gender Other (please specify)	390	0.0605	0.262	0.231	0.817	-0.453	0.574
Race Am Ind or AK Native	162	-0.1885	0.191	-0.989	0.323	-0.562	0.185
Race Asian	3755	0.1146	0.08	1.433	0.152	-0.042	0.271
Race Black or African American	464	-0.1848	0.129	-1.435	0.151	-0.437	0.068
Race White	8215	0.1199	0.08	1.507	0.132	-0.036	0.276
Education Associate degree	80	-0.0003	†	-2.39E-10	1	-‡	‡
Education Bachelor's degree	4232	-0.0621	†	-4.81E-08	1	-‡	‡
Education Doctorate	1162	-0.129	†	-1E-07	1	-‡	‡
Education HS deg. or equiv.	384	0.0031	†	2.41E-09	1	-‡	‡
Education Master's degree	4198	-0.1083	†	-8.4E-08	1	-‡	‡
Education Other (please specify)	306	-0.2139	†	-1.66E-07	1	-‡	‡
Education Professional degree	668	-0.2233	†	-1.73E-07	1	-‡	‡
Education Some college no deg.	1867	-0.1269	†	-9.84E-08	1	-‡	‡
Language English	9228	0.0743	0.19	0.391	0.695	-0.298	0.446
Language Other (please specify)	3513	0.0099	0.189	0.052	0.958	-0.361	0.381
BalacingElements diff	n/a	0.0635	0.095	0.669	0.504	-0.123	0.25
ColorHarmony diff	n/a	0.3351	0.073	4.61	0	0.193	0.478
Content diff	n/a	0.3625	0.044	8.294	0	0.277	0.448
DoF diff	n/a	-0.0275	0.068	-0.405	0.685	-0.161	0.106
Light diff	n/a	0.0567	0.059	0.957	0.339	-0.059	0.173
MotionBlur diff	n/a	0.2912	0.149	1.954	0.051	-0.001	0.583
Object diff	n/a	0.032	0.045	0.717	0.473	-0.056	0.12
Repetition diff	n/a	0.0535	0.119	0.45	0.652	-0.179	0.287
RuleOfThirds diff	n/a	0.0069	0.085	0.08	0.936	-0.161	0.174
Symmetry diff	n/a	0.2163	0.169	1.281	0.2	-0.115	0.547
VividColor diff	n/a	-0.0229	0.057	-0.402	0.688	-0.135	0.089

Table 1: Regression coefficients part 1. Above, † indicates a value of 129000 and ‡ indicates a value of 2530000. These large error margins are caused by close-to collinear features for Education. "const" is the intercept term, and its N value is the total number of image comparisons in the dataset.

Table 3 (left) shows the confusion matrix between labels inferred from the AADB image scores and the PR-AADB labels. (right) shows the confusion matrix between labels inferred from scores inferred from the PR-AADB labels. There are miss-classifications here because no single set of scores can predict many different participants' choices.

Table 4 (left) shows the confusion matrix between labels predicted using a deep classifier trained on the Flickr-AES dataset. (right) shows those results after fine-tuning using a SVM, which predicts based on content and aesthetic attributes, in addition to the raw label. The very slight increase in accuracy is the result of more pairs correctly classified as 0, which is counterbalanced by more pairs with PR-AADB labels of 1 and -1 getting misclassified.

Feature	**N**	**coef**	**std err**	**z**	**P > |z|**	**0.025**	**0.975**
American egret	109	0.5594	0.198	2.832	0.005	0.172	0.947
limpkin	77	0.556	0.243	2.286	0.022	0.079	1.033
brown bear	26	1.865	0.549	3.399	0.001	0.789	2.941
dragonfly	49	0.8541	0.3	2.843	0.004	0.265	1.443
monarch	45	0.8112	0.33	2.46	0.014	0.165	1.458
baseball	315	-0.0926	0.122	-0.761	0.447	-0.331	0.146
breakwater	340	-0.2677	0.116	-2.302	0.021	-0.496	-0.04
cab	325	-0.1789	0.12	-1.495	0.135	-0.414	0.056
canoe	331	0.1545	0.114	1.357	0.175	-0.069	0.378
carousel	99	-0.3873	0.22	-1.762	0.078	-0.818	0.044
church	442	0.1962	0.105	1.867	0.062	-0.01	0.402
cloak	106	-0.4319	0.215	-2.012	0.044	-0.853	-0.011
cowboy hat	189	-0.1544	0.156	-0.987	0.324	-0.461	0.152
crutch	497	-0.186	0.097	-1.922	0.055	-0.376	0.004
dock	405	0.1611	0.104	1.542	0.123	-0.044	0.366
flagpole	174	-0.2544	0.16	-1.588	0.112	-0.568	0.06
football helmet	326	-0.2734	0.145	-1.89	0.059	-0.557	0.01
fountain	361	0.3	0.109	2.753	0.006	0.086	0.514
gown	165	-0.1772	0.168	-1.052	0.293	-0.507	0.153
greenhouse	383	0.1882	0.107	1.766	0.077	-0.021	0.397
grille	116	-0.455	0.205	-2.218	0.027	-0.857	-0.053
harmonica	116	-0.3385	0.199	-1.702	0.089	-0.728	0.051
kimono	327	-0.301	0.122	-2.475	0.013	-0.539	-0.063
limousine	195	0.3608	0.148	2.432	0.015	0.07	0.652
mask	87	-0.5434	0.241	-2.257	0.024	-1.015	-0.071
military uniform	273	-0.4184	0.135	-3.096	0.002	-0.683	-0.154
minibus	264	0.1552	0.128	1.216	0.224	-0.095	0.405
miniskirt	190	-0.192	0.156	-1.231	0.218	-0.498	0.114
monastery	546	-0.1246	0.097	-1.288	0.198	-0.314	0.065
parachute	118	0.5243	0.19	2.757	0.006	0.152	0.897
park bench	283	-0.1546	0.125	-1.234	0.217	-0.4	0.091
ping-pong ball	121	0.3859	0.187	2.067	0.039	0.02	0.752
quill	74	0.5	0.245	2.041	0.041	0.02	0.98
restaurant	592	0.1083	0.086	1.253	0.21	-0.061	0.278
rifle	73	-0.6336	0.272	-2.332	0.02	-1.166	-0.101
rugby ball	342	-0.2781	0.14	-1.987	0.047	-0.552	-0.004
sarong	257	-0.0727	0.135	-0.54	0.589	-0.336	0.191
snorkel	118	0.479	0.19	2.525	0.012	0.107	0.851
stone wall	362	0.1562	0.109	1.434	0.152	-0.057	0.37
suspension bridge	309	0.1894	0.117	1.613	0.107	-0.041	0.42
swimming trunks	152	-0.2801	0.174	-1.61	0.107	-0.621	0.061
unicycle	512	-0.1622	0.095	-1.699	0.089	-0.349	0.025
vault	333	-0.2449	0.119	-2.049	0.04	-0.479	-0.011
vestment	328	0.2176	0.115	1.886	0.059	-0.009	0.444
water tower	109	-0.544	0.209	-2.601	0.009	-0.954	-0.134
wig	190	-0.1675	0.154	-1.085	0.278	-0.47	0.135
window screen	108	-0.6873	0.218	-3.154	0.002	-1.114	-0.26
wreck	118	0.4598	0.189	2.433	0.015	0.089	0.83
crossword puzzle	86	0.4892	0.225	2.175	0.03	0.048	0.93
promontory	434	0.2427	0.101	2.413	0.016	0.046	0.44
daisy	180	0.254	0.154	1.654	0.098	-0.047	0.555

Table 2: Regression coefficients part 2.

PR-AADB labels

Labels inferred from AADB

	-1	0	1
-1	2301	936	1120
0	1672	1055	1732
1	1164	988	2324

Labels inferred from PR-AADB

	-1	0	1
-1	3085	1105	167
0	1058	2313	1088
1	168	990	3318

Table 3: Confusion Matrix across all participants for Figure 4 (Center) in the main paper. The left table indicates X values (i.e. labels inferred from the AADB single-image scores) and the right table indicates Y values (i.e. labels inferred from the scores, which were in turn inferred from the PR-AADB labels).

PR-AADB labels

Raw predictions

	-1	0	1
-1	1399	2122	955
0	905	2193	1361
1	563	1898	1896

Finetuned predictions

	-1	0	1
-1	1188	2203	1085
0	660	3134	665
1	1096	2028	1233

Table 4: Confusion Matrix across all participants for Figure 4 (Right) in the main paper. The left table indicates X values (i.e. raw labels predicted by a deep CNN classifier) and the right table indicates Y values (i.e. labels finetuned from the raw labels using a SVM classifier).