Supplementary Material: Understanding and Mitigating Annotation Bias in Facial Expression Recognition

Yunliang Chen Jungseock Joo
University of California, Los Angeles
chenyunliang@ucla.edu, jjoo@comm.ucla.edu

1. Angry Annotation Bias Between Genders

1.1. OpenFace Accuracy for Angry AUs

We evaluate annotation bias for the angry expression. Recall that anger is defined as the combination of AU4 (brow lowerer), AU5 (upper lid raiser), AU7 (lid tightener), and AU23 (lip tightener). In this section, we check the quality of OpenFace AU recognition by comparing its output with the expert-coded EmotioNet dataset that consists of 24,600 images [3]. Since EmotioNet does not have labels for AU7 and AU23, we are only able to check for AU4 and AU5.

Table 1 shows the OpenFace AU accuracy results for AU4 and AU5. Similar to the case of happiness expression, we binarize the AU intensity outputs of OpenFace using a threshold that is chosen to maximize the overall accuracy of prediction. The accuracy result is shown as "Raw" in Table 1. The difference between males and females is small for AU5, while it is not negligible for AU4. Thus, we re-calibrate the OpenFace output so that different binarization thresholds are chosen to balance the accuracy between males and females. We remark that learning subgroup-specific thresholds is a common technique used to achieve fairness [12]. As shown in the "Recalibrated" column of Table 1, there is no statistically significant difference between males and females after calibration (*i.e.*, the AU's are fair). Since there is little difference in AU5 before and after calibration, we will only apply calibration to AU4 in the below evaluations.

	AU4 Accuracy		AU5 Accuracy	
	Raw	Recalibrated	Raw	Recalibrated
Male	0.854	0.857	0.968	0.958
Female	0.928	0.855	0.958	0.958
p-value	0.000	0.779	0.00002	0.880

Table 1: Accuracy of OpenFace AU Recognition, evaluated on 24,600 EmotioNet images with expert-coded AUs. For the raw accuracy, the AU intensity output of OpenFace are binarized using a single threshold that maximizes the overall accuracy, while for the re-calibrated accuracy, different binarization thresholds are chosen to balance the accuracy between males and females.

1.2. Annotation Bias of the Angry Expression

In this section, we present evaluation results for the angry annotation bias in major public datasets. Similar to the evaluation of happy annotations, we apply the OpenFace AU detector and obtain the AU presence and AU intensity information for each image. AU4 intensities are then binarized into AU4 presence variable using the calibrated thresholds found earlier. All other AUs (AU5, AU7, and AU23) use the raw AU presence outputs from OpenFace since no adjustment is needed or available as shown in the previous section. We also apply our gender classifier when the gender information is not available (*i.e.*, for ExpW and AffectNet).

Table 2 shows the proportion of "angry" labels among males and females conditioned on different values of AU4, AU5, AU7, and AU23. For each of the conditional distributions of "angry," a chi-square test of independence is used to determine whether there is a significant relationship between the labels and gender after controlling for the AUs. Unlike Table 2 in the paper, only marginal distributions are shown since the joint distribution of (AU4, AU5, AU7, AU23) can take $2^4 = 16$ possible values and in many cases do not contain enough data or significantly reduce the power of statistical testing. This is expected since many AU's are correlated and so some combinations of (AU4, AU5, AU7, AU23) are much more likely than others. We believe that the distributions of "angry" labels conditioned on marginal AU still provides useful information for comparing the labels of lab-controlled datasets and in-the-wild datasets.

From Table 2, we can see significant differences between lab-controlled datasets and in-the-wild datasets. The pattern is similar to that of the happiness expression. For both KDEF and CFD, the distribution of "angry" labels is independent of gender when the AUs are controlled. On the other hand, for ExpW, RAF-DB, and AffectNet, the proportion of "angry" labels is significantly higher for males than

Data (Collecting Condition, Size)	Conditioned on Marginal AU					Conditioned on Marginal AU				
	AU	P(Angry |AU, M)	P(Angry |AU, F)	Δ	p-value of χ^2 test for $Y \perp\!\!\!\perp Z$	AU	P(Angry |AU, M)	P(Angry |AU, F)	Δ	p-value of χ^2 test for $Y \perp\!\!\!\perp Z$
KDEF (Lab, 980) [8]	AU4=0	0.090	0.060	0.030	0.103	AU7=0	0.124	0.089	0.035	0.246
	AU4=1	0.405	0.465	-0.061	0.408	AU7=1	0.162	0.178	-0.016	0.609
	AU5=0	0.207	0.173	0.034	0.320	AU23=0	0.123	0.133	-0.009	0.684
	AU5=1	0.071	0.104	-0.032	0.231	AU23=1	0.268	0.220	0.047	0.533
CFD (Lab, 1,207) [9]	AU4=0	0.075	0.091	-0.016	0.324	AU7=0	0.026	0.043	-0.017	0.228
	AU4=1	1.000	0.933	0.067	-	AU7=1	0.245	0.233	0.012	0.743
	AU5=0	0.132	0.156	-0.023	0.326	AU23=0	0.081	0.074	0.007	0.711
	AU5=1	0.096	0.072	0.024	0.448	AU23=1	0.208	0.237	-0.028	0.490
ExpW (Web, 91,793) [15, 16]	AU4=0	0.041	0.033	0.008	0.000 ***	AU7=0	0.034	0.026	0.008	0.000 ***
	AU4=1	0.145	0.126	0.019	0.235	AU7=1	0.053	0.044	0.008	0.000 ***
	AU5=0	0.042	0.035	0.007	0.000 ***	AU23=0	0.045	0.036	0.009	0.000 ***
	AU5=1	0.046	0.035	0.010	0.000 ***	AU23=1	0.041	0.033	0.008	0.010 **
RAF-DB (Web, 15,339) [7]	AU4=0	0.089	0.030	0.058	0.000 ***	AU7=0	0.052	0.017	0.036	0.000 ***
	AU4=1	0.295	0.061	0.234	0.000 ***	AU7=1	0.127	0.046	0.081	0.000 ***
	AU5=0	0.108	0.037	0.071	0.000 ***	AU23=0	0.089	0.031	.058	0.000 ***
	AU5=1	0.043	0.013	0.030	0.000 ***	AU23=1	0.097	0.032	.066	0.000 ***
AffectNet-Manual (Web, 420,299) [10]	AU4=0	0.083	0.036	0.047	0.000 ***	AU7=0	0.088	0.037	0.051	0.000 ***
	AU4=1	0.287	0.112	0.175	0.000 ***	AU7=1	0.093	0.039	0.054	0.000 ***
	AU5=0	0.093	0.037	0.056	0.000 ***	AU23=0	0.092	0.037	0.054	0.000 ***
	AU5=1	0.084	0.038	0.046	0.000 ***	AU23=1	0.086	0.038	0.048	0.000 ***
AffectNet-Automatic (Web, 539,607) [10]	AU4=0	0.095	0.019	0.066	0.000 ***	AU7=0	0.081	0.017	0.064	0.000 ***
	AU4=1	0.374	0.118	0.256	0.000 ***	AU7=1	0.108	0.026	0.083	0.000 ***
	AU5=0	0.103	0.023	0.080	0.000 ***	AU23=0	0.096	0.019	0.077	0.000 ***
	AU5=1	0.071	0.018	0.053	0.000 ***	AU23=1	0.085	0.029	0.056	0.000 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 2: Proportion of "angry" labels among males and females conditioned on AU4, AU5, AU7, and AU23 for each of the popular expression datasets. Here $Y \in \{0, 1\}$ is the "angry" label, $Z \in \{M, F\}$ is the gender attribute. Blanks and omitted p-values indicate that the AUs do not contain enough data for the chi-square tests.

females even when the AUs have been controlled.

Figure 1 shows the proportion of "angry" labels as a function of AU intensities for each in-the-wild dataset. The proportion of "angry" labels is higher when the AU intensities are higher, but the effect is different between males and females. All in-the-wild datasets show large discrepancies in the conditional distributions of "angry" labels between males and females. This is consistent with the result in Table 2, and we conclude that angry annotation bias is a prominent issue for in-the-wild datasets.

1.3. Bias Correction for the Angry Expression

In this section, we examine the effectiveness of the proposed AUC-FER algorithm in removing angry annotation bias. Similar to the experiments we did for the happiness expression, we compare our algorithm with other debiasing methods in the fairness literature, including uniform confusion [1], gradient reversal [14], domain discriminative training [13], and domain independent training [13].

To test for robustness of our algorithm, we use AffectNet-Automatic as our training data instead of ExpW as we did for the happiness expression evaluation. The training data is of size 20,000 and is randomly sampled from AffectNet-Automatic. The test set is still constructed from CFD [9]. In particular, we remove from CFD a few easy non-angry images (whose predicted scores from a pre-trained naive classifier are less than 0.05) and then balance the number of male and female images in each AU combination. Doing so also balances the number of angry images between males and females. Similar to the experiments for the happiness expression, the thresholds for binarizing the output of the trained classifier are adjusted to maximize the accuracy on the test set.

For all experiments, we use the ResNet-50 architecture [4] pre-trained on ImageNet in PyTorch. The baseline model is a naive classifier fine-tuned by Adam optimization [6] with a learning rate of 0.0001 in PyTorch. For the four benchmark models, we follow Wang *et al*. [13] and replace the FC layer of the ResNet-50 model with two consecutive FC layers both of size 2048 with Dropout and ReLU in between. For AUC-FER, we use the PyTorch Metric Learning library [11] for the triplet loss implementation, and the hyperparameter λ which trades off $\mathcal{L}_{softmax}$ and $\lambda\mathcal{L}_{trp}$ is set to 10. The training data is resampled so that the number of

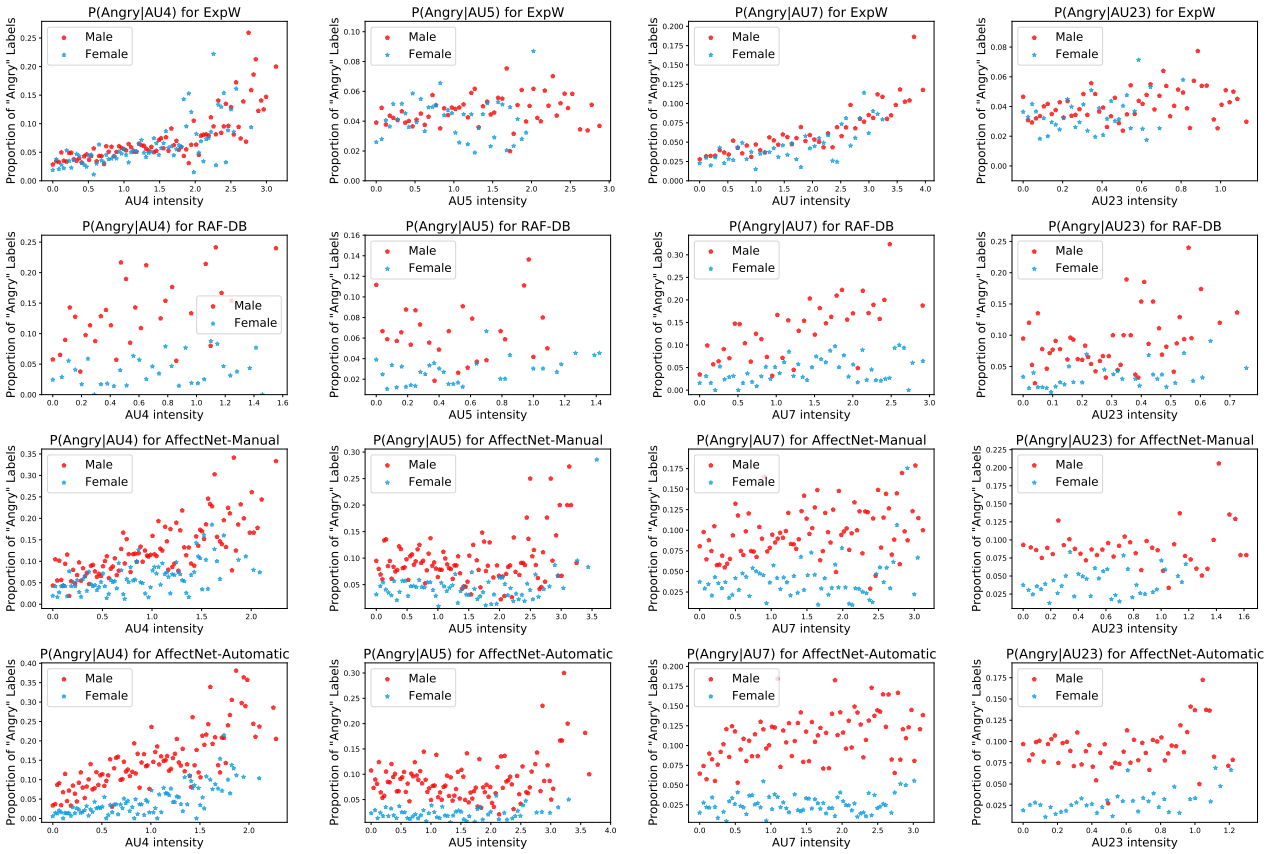

Figure 1. Proportion of "angry" labels among males and females conditioned on AU4, AU5, AU7, and AU23 intensity for each in-the-wild expression dataset. Significant differences can be seen between males and females, indicating the presence of annotation bias.

images is balanced for each gender and AU combination. The experiment is repeated 5 times for each model.

Since the attribute we are interested in is "angry", the Calders-Verwer (CV) discrimination score [2] here is defined as:

$$Disc = |P(\hat{Y} = Angry|M) - P(\hat{Y} = Angry|F)|. \quad (1)$$

Table 3 shows the discrimination scores for all debiasing methods and compares them against the baseline model. All methods achieve a reduction in bias, but AUC-FER obtains the lowest discrimination score, which is a 72% reduction in bias compared to the baseline model. This again shows that the proposed AUC-FER algorithm is effective in removing annotation bias.

Methods (ResNet-50 [4])	*Disc*	Compared to Baseline (%)
Baseline	0.064 ± 0.020	-
Uniform Confusion [1]	0.044 ± 0.025	68.3
Gradient Projection [14]	0.031 ± 0.031	49.2
Domain Discriminative [13]	0.041 ± 0.051	63.5
Domain Independent [13]	0.021 ± 0.018	33.3
AUC-FER (Ours)	$\mathbf{0.018 \pm 0.039}$	**28.2**

Table 3: Discrimination scores for various debiasing methods using the ResNet-50 architecture trained on random subsets of AffectNet-Automatic of size 20,000 and tested on CFD for the anger expression.

2. Happy Annotation Bias Across Age Groups

In this section, we describe the evaluation of annotation bias for the happy expression across age groups. We note that most of the datasets (both lab-controlled and in-the-wild) are severely dominated by younger people, making the statistical tests challenging.

Since only RAF-DB includes age labels, we first train a simple age classifier with ResNet-34 architecture [4] using the FairFace dataset [5] similar to that for training a gender classifier. We then apply the trained classifier on the datasets that do not have age labels (*i.e.*, KDEF, CFD, ExpW, and AffectNet). The age predictions are then grouped into the following 4 groups: "less than 19", "20-39", "40-59", and "more than 60." The original labels provided by RAF-DB are "0-3," "4-19," "20-39," "40-69," and "70+." We group the "0-3" and "4-19" age groups together to increase the number of samples in each age group. In

Data	Metrics P(Happy|AU6, AU12)	4 Age Groups					2 Age Groups		
		≤ 19	20-39	40-59[1]	≥ 60[1]	p-value	≤ 39	≥ 40	p-value
ExpW [15, 16]	P(Happy|(0, 0))	0.169	**0.190**	0.177	0.207	0.000 ***	0.187	0.191	0.382
	P(Happy|(1, 0))	0.190	0.262	0.251	**0.303**	0.017 *	0.251	0.274	0.306
	P(Happy|(0, 1))	**0.763**	0.708	0.586	0.626	0.000 ***	**0.716**	0.606	0.000 ***
	P(Happy|(1, 1))	**0.838**	0.832	0.765	0.806	0.000 ***	**0.833**	0.785	0.000 ***
	P(Happy)	0.321	0.335	0.299	0.346		0.333	0.320	
	Number of samples	11,561	62,622	6,033	5,139		74,183	11,172	
RAF-DB [7]	P(Happy|(0, 0))	0.231	0.168	**0.249**	0.130	0.000 ***	**0.234**	0.168	0.000 ***
	P(Happy|(1, 0))	**0.320**	0.180	0.285	0.250	0.003 **	**0.410**	0.180	0.000 ***
	P(Happy|(0, 1))	0.859	0.837	0.891	0.741	0.209	0.877	0.837	0.076 .
	P(Happy|(1, 1))	0.837	0.857	**0.903**	0.879	0.010 **	**0.903**	0.857	0.048 *
	P(Happy)	0.379	0.340	0.536	0.433		0.438	0.340	
	Number of samples	3,205	6,805	1,911	293		10,111	2,103	
AffectNet-Manual [10]	P(Happy|(0, 0))	0.141	**0.157**	0.078	0.078	0.000 ***	**0.112**	0.078	0.000 ***
	P(Happy|(1, 0))	**0.407**	0.241	0.225	0.272	0.023 *	0.272	0.240	0.287
	P(Happy|(0, 1))	**0.746**	0.697	0.599	0.525	0.000 ***	**0.705**	0.584	0.000 ***
	P(Happy|(1, 1))	0.824	**0.839**	0.797	0.749	0.000 ***	**0.837**	0.784	0.000 ***
	P(Happy)	0.342	0.324	0.282	0.282		0.327	0.282	
	Number of samples	4,041	20,486	8,477	2,690		24,527	11,167	
AffectNet-Automatic [10]	P(Happy|(0, 0))	**0.244**	0.184	0.160	0.153	0.000 ***	**0.198**	0.158	0.000 ***
	P(Happy|(1, 0))	0.591	0.428	0.435	**0.436**	0.017 *	0.478	0.436	0.210
	P(Happy|(0, 1))	**0.867**	0.858	0.822	0.711	0.000 ***	**0.860**	0.800	0.000 ***
	P(Happy|(1, 1))	0.887	0.932	**0.932**	0.882	0.000 ***	0.924	0.918	0.436
	P(Happy)	0.437	0.423	0.394	0.379		0.426	0.390	
	Number of samples	6,876	24,111	8,432	2,865		30,987	11,297	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[1] For RAF-DB, the 4 age groups are ≤ 19, 20 - 39, 40 - 69, and ≥ 70.

Table 4: Conditional and marginal distributions of "happy" labels along with the numbers of samples for each age group for each in-the-wild dataset. The p-values are the χ^2 tests for independence of the "happy" labels and the age groups. When the p-values are significant at the 0.05 level, the age groups with the highest proportion of "happy" labels are highlighted.

addition to this 4-group categorization, we also group the images into "less than 40" and "more than 40" to further alleviate the problem of having too few samples in some age groups when conditioning on the AUs and to investigate general discrepancies between younger and older people. However, the lab-controlled datasets contain too few images in the older age groups (*e.g.*, only 26 out of the 1,207 total images are predicted to be older than 40 years old). As a result, we are unable to evaluate the annotation bias for the two lab-controlled datasets.

Table 4 shows the conditional and marginal distributions of "happy" labels along with the numbers of samples for each age group for each in-the-wild expression dataset. We see that all datasets are heavily dominated by younger people. The sum of the numbers for each age group does not add up to those of the full datasets due to the fact that OpenFace fails to produce AU labels (possibly due to occlusion or blur) for a small fraction of the images. The p-values are the χ^2 tests for independence of the "happy" labels and the age groups. When the p-values are significant at the 0.05 level, the age groups with the highest proportion of "happy" labels are highlighted. We can see that the differences in the proportion of "happy" labels are statistically significant in most cases. Moreover, younger age groups have a higher proportion of "happy" labels in general, suggesting the existence of systematic annotation bias.

Figure 2 plots the annotation bias of the "happy" expression across the 4 age groups. The first row shows the proportions of "happy" labels for each of the 4 age groups conditioned on AU6 and AU12 presence variables. The error bars indicate one standard error of the proportion. The second and third rows show the fitted logistic regression curves as a function of AU intensities. The shaded regions indicate 95% confidence intervals. From the plots, we see that AU6 exhibits a larger bias than AU12. Figure 3 compares younger and older populations. Consistent with Table 4, we see that younger people are more likely to be annotated as "happy" compared to older people, although the saliency varies across datasets. Further analysis on more balanced datasets (*i.e.*, datasets that have more older-than-40 and especially older-than-60 people) is needed.

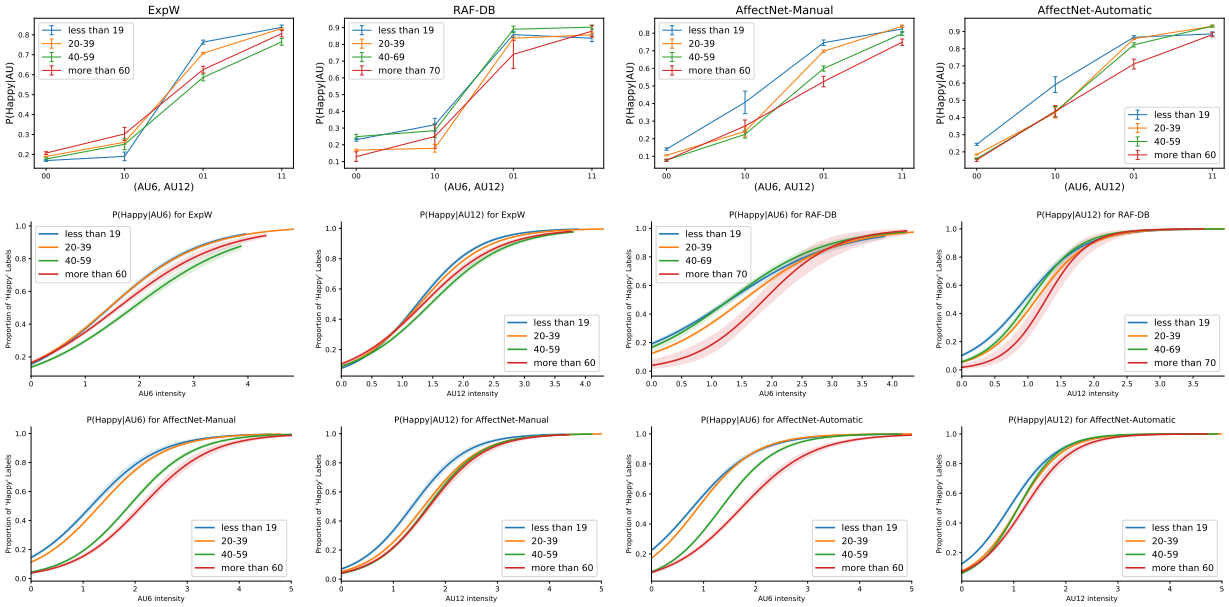

Figure 2. Annotation bias of the "happy" expression across 4 age groups for each in-the-wild expression dataset. The first row shows the proportions of "happy" labels for each of the 4 age groups conditioned on AU6 and AU12 presence variables. The error bars indicate one standard error of the proportion. The second and third rows show the fitted logistic regression curves as a function of AU intensities. The shaded regions indicate 95% confidence intervals. AU6 exhibits a larger bias than AU12.

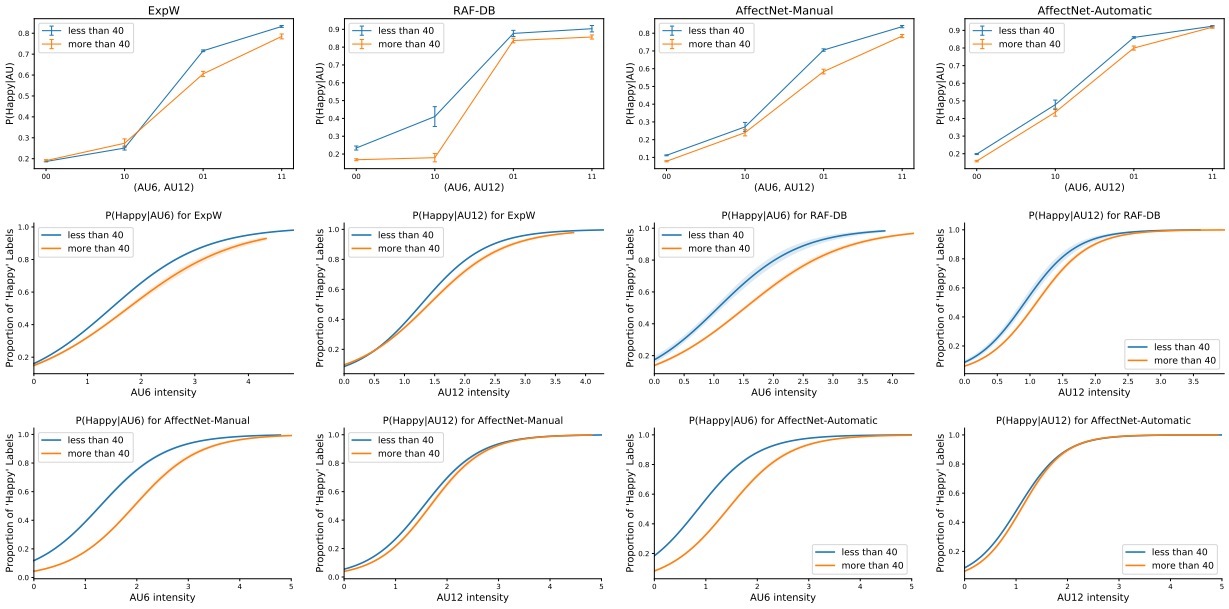

Figure 3. Annotation bias of the "happy" expression between younger and older populations. The first row shows the proportions of "happy" labels conditioned on AU6 and AU12 presence variables. The error bars indicate one standard error of the proportion. The second and third rows show the fitted logistic regression curves as a function of AU intensities. The shaded regions indicate 95% confidence intervals. Younger people seems more likely to be annotated as "happy" compared to older people.

3. Happy Annotation Bias Across Racial Groups

In this section, we describe the evaluation of annotation bias for the happy expression across racial groups. Similar to the issue of imbalanced classes we encounter with age groups, many of the in-the-wild datasets (with the exception of ExpW) are severely dominated by white people, making the statistical tests challenging. Again, lab-controlled

Data	Metrics P(Happy|AU6, AU12)	3 Groups				6 Groups			
		Asian	Black	White	p-value	Indian	Latino-Hispanic	Middle Eastern	p-value
ExpW [15, 16]	P(Happy|(0, 0))	0.177	0.181	**0.194**	0.000 ***	0.242	0.21	0.177	0.000 ***
	P(Happy|(0, 1))	0.697	**0.723**	0.678	0.004 **	0.75	0.763	0.685	0.000 ***
	P(Happy|(1, 0))	0.253	0.267	0.243	0.67	0.327	0.275	0.235	0.654
	P(Happy|(1, 1))	0.815	0.833	0.81	0.252	0.882	0.88	0.798	0.000 ***
	P(Happy)	0.319	0.323	0.328		0.406	0.398	0.302	
	Number of samples	31,791	11,381	25,826		1,701	8,751	5,905	
RAF-DB [7]	P(Happy|(0, 0))	0.2	**0.249**	0.187	0.000 ***				
	P(Happy|(0, 1))	0.883	0.833	0.847	0.286				
	P(Happy|(1, 0))	**0.343**	0.244	0.216	0.023 *				
	P(Happy|(1, 1))	0.873	0.919	0.862	0.177				
	P(Happy)	0.403	0.421	0.375					
	Number of samples	1,904	1,011	9,299					
AffectNet-Manual [10]	P(Happy|(0, 0))	0.094	0.083	**0.106**	0.005 **	0.099	0.104	0.083	0.003 **
	P(Happy|(0, 1))	0.657	0.693	0.683	0.499	0.68	0.68	0.627	0.228
	P(Happy|(1, 0))	0.208	0.275	0.266	0.663	0.229	0.212	0.19	0.654
	P(Happy|(1, 1))	0.781	0.838	0.821	0.083 .	0.817	0.826	0.718	0.004 **
	P(Happy)	0.311	0.320	0.320		0.312	0.321	0.222	
	Number of samples	2,211	2,833	23,780		1,159	3,179	2,532	
AffectNet-Automatic [10]	P(Happy|(0, 0))	**0.207**	0.154	0.193	0.000 ***	0.176	0.204	0.153	0.000 ***
	P(Happy|(0, 1))	0.828	0.847	0.853	0.254	0.837	0.885	0.784	0.002 **
	P(Happy|(1, 0))	0.436	0.435	0.473	0.709	0.262	0.483	0.413	0.164
	P(Happy|(1, 1))	0.874	0.918	**0.930**	0.000 ***	0.917	0.932	0.874	0.000 ***
	P(Happy)	0.427	0.391	0.430		0.403	0.453	0.281	
	Number of samples	3,619	4,454	26,798		1,548	3,198	2,667	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[1] The provided labels from RAF-DB are "Asian," "White," and "Black" only.

Table 5: Conditional and marginal distributions of "happy" labels along with the numbers of samples for each racial/ethnic group for each in-the-wild dataset. The p-values are the χ^2 tests for independence of the "happy" labels and the racial/ethnic groups. When the p-values for the 3 racial groups are significant at the 0.05 level, the racial groups with the highest proportion of "happy" labels are highlighted.

datasets are too small and lack diversity in their populations (*e.g.*, CFD has only 100 images for Asian and Latino groups each), and so we only focus on in-the-wild datasets.

Since only RAF-DB includes race labels, we first train a simple race classifier with ResNet-34 architecture [4] using the FairFace dataset [5] similar to the procedure for training the gender and age classifiers. We follow the convention of the race/ethnicity categorization of FairFace, which classifies each image into one of seven groups: White, Black, Latino/Hispanic, East Asian, Southeast Asian, Indian, and Middle Eastern. We then apply the trained classifier on the datasets that do not have age labels (*i.e.*, KDEF, CFD, ExpW, and AffectNet). We combine the East Asian and Southeast Asian groups together. This results in 6 racial/ethnic groups. However, since the race labels provided by RAF-DB only include White, Black, and Asian, and many datasets contain relatively few images for Indian, Middle Eastern, and Latino/Hispanic groups, we also conduct analysis that only focuses on the three major races.

Table 5 shows the conditional and marginal distributions of "happy" labels along with the numbers of samples for each racial/ethnic group for each in-the-wild expression dataset. As mentioned in the previous section, the sum of the numbers for each racial group does not add up to those of the full datasets due to the fact that OpenFace fails to produce AU labels (possibly due to occlusion or blur) for a small fraction of the images. The p-values are the χ^2 tests for independence of the "happy" labels and the racial/ethnic groups. When the p-values for the 3 racial groups are significant at the 0.05 level, the racial groups with the highest proportion of "happy" labels are highlighted. We can see that even though the differences in the proportion of "happy" labels are sometimes statistically significant, the bias is mostly idiosyncratic. In other words, there is no consistent pattern of systematic annotation bias that one group is more or less likely to be annotated "happy" than others.

Figure 4 plots the annotation bias of the "happy" expression across the racial and ethnic groups. The first row shows the proportions of "happy" labels for each of the three major racial groups only, while the second row plots all six

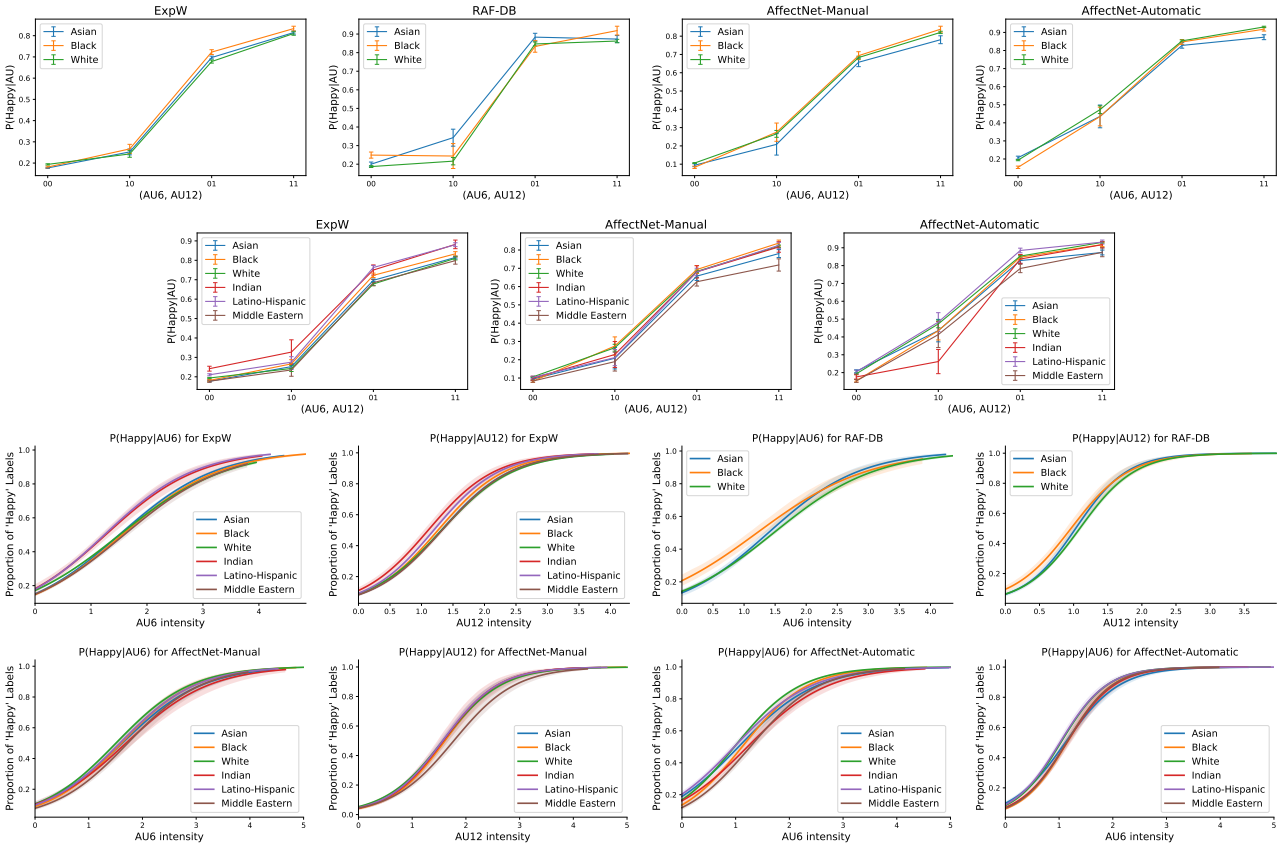

Figure 4. Annotation bias of the "happy" expression across racial/ethnic groups for each in-the-wild expression dataset. The first row shows the proportions of "happy" labels for each of the three major racial groups only while the second row plots all six groups. The error bars indicate one standard error of the proportion. The last two rows show the fitted logistic regression curves as a function of AU intensities. 95% confidence intervals are indicated by shaded regions.

groups. The last two rows show the fitted logistic regression curves as a function of AU intensities. Consistent with the results in Table 5, we see that the differences among the racial and ethnic groups are minor, and no consistent bias exists across all datasets. Further analysis on other expressions would ideally require more balanced datasets (*i.e.*, datasets that have more minority races).

References

[1] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[2] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[3] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5562–5570, 2016.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[7] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 2584–2593. IEEE, 2017.

[8] Daniel Lundqvist, Anders Flykt, and Arne Öhman. The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, 91(630):2–2, 1998.

[9] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, 47(4):1122–1135, 2015.

[10] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, PP(99):1–1, 2017.

[11] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. Pytorch metric learning, 2020.

[12] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–1, 2020.

[13] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8919–8928, 2020.

[14] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[15] Zhanpeng Zhang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Learning social relation traits from face images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3631–3639, 2015.

[16] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018.