Supplemental Material for Multi-Dimensional, Nuanced and Subjective – Measuring the Perception of Facial Expression

De'Aira Bryant[+,‡ *] Siqi Deng[+] Nashlie Sephus[+] Wei Xia[+ †] Pietro Perona[+,‡‡]
+ AWS AI Labs, ‡ Georgia Institute of Technology, ‡‡ California Institute of Technology
dbryant@gatech.edu, {siqideng, nashlies, peronapp}@amazon.com, weixiaee@gmail.com

Appendix A. Annotation collection

Amazon SageMaker GroudTruth is used to collect multi-dimensional intensity ratings from Mechanical Turk workers. A custom annotator interface was developed to obtain ratings on the 1000-image set for the 6 primary expressions (Fig. 1), the 15 compound expressions (Fig. 2) and the set of 21 primary and compound expressions (Fig. 3). A rating was required for each expression dimension during each experiment. Annotators were paid for each image with a complete set of intensity ratings. Sample images from the 1000-set are shown in Figure 10 and Figure 11 with the original ExpW label, a histogram of normalized annotator intensity ratings, and the resulting beta distribution per emotion per face.

We further examine annotator behavior during the compound and 21 expression annotation experiments. Fig. 4 illustrates the log_{10} time in seconds for annotators to complete a single annotation task for 6 primary, 15 compound, or 21 (primary + compound) expressions during each experiment. The frequency of each intensity ratings across all images and annotators per expression dimension can be seen in Figures 5 and 6 for the compound and 21 expressions respectively. As observed in the primary expression experiment, the range of the scale utilized by annotators often differs by expression dimension. When providing annotations along 21 dimensions, annotators utilized the middle of the intensity scale less than those in the primary and compound experiments. This suggests that as the number of expression dimensions increase, annotators are less likely to indicate moderate interpretations of facial expressions (Sec. 4.2).

We also consider the histograms of entropy scores across all images annotated for the compound and 21 expressions as displayed in Figures 7 and 8 (entropy (S) as defined in Eq. 1, Sec. 3.3) Lower entropy scores signify greater agreement between annotators. Notably, annotators in the 21 expressions experiment agreed less on their ratings of primary expressions than those who annotated primary expressions exclusively (See Fig. 8 in main paper and Fig. 8). Overall, variance in annotator agreement increases as the number of expression dimensions increases.

Appendix B. Data conditioning and modeling

Before fitting the Beta distribution, as a pre-processing step, we normalized the annotator ratings as described in the main paper (Sec. 3.2). However, other annotator intensity rating preprocessing methods can be considered to effectively estimate the expression distributions. We explore this possibility by comparing 4 techniques to transform raw annotator ratings prior to estimating the parameters of a beta distribution: normalization (used for all experiments in main paper), baseline, weighted, and hybrid (See Figure 9).

The discrete raw intensity ratings of the annotators take values $v_{(i,d,l)} \in \{0, 1, 2, 3, 4\}$ where 0 indicates 'Not at All' and 4 indicates 'Extremely'. The normalization method transforms each raw rating $v_{(i,d,l)}$ into a normalized rating $r_{(i,d,l)}$ as defined in Section 3.2 of the paper. This method was used to conduct all experiments in the paper.

Next, we consider that any expression may be perceived by an observer at any intensity. The baseline method appends the set of normalized annotator ratings $r_{i,d}$ with the set $\{$ 0.1, 0.3, 0.5, 0.7, 0.9 $\}$. The total number of intensity ratings used by the Baseline method to estimate the parameters of a distribution per expression d for a face image i is $L + 5$ where L is the total number of annotator ratings per image and typically $L = 9$. Figure 9 illustrates how the baseline method initializes each possible intensity bin.

We then consider that an observer's true perception of an expression may actually fall between the possible discrete intensity values. The weighted method transforms each normalized rating into 4 ratings using a weighted scale where $W(r_{(i,d,l)}) = [r_{(i,d,l)} - 0.1, r_{(i,d,l)}, r_{(i,d,l)}, r_{(i,d,l)} + 0.1)]$. Thus, the weighted method utilizes 10 bins for ratings and the total number of ratings used to estimate the shape parameters per expression d for a face image i is $4L$. Figure 9

*Work done during an Amazon internship.
†Work done when at Amazon.

illustrates how the weighted method transforms an intensity rating.

The hybrid method combines the features of the baseline and weighted method such that each bin in initialized with a single intensity rating (i.e. the set of normalized annotator ratings $r_{i,d}$ is appended with the set $\{$ 0.0, 0.1, ... 1.0 $\}$. and annotator ratings are then transformed using $W(r_{(i,d,l)})$. Therefore, the total number of ratings used to estimate the parameters using the hybrid method is $4L + 10$. Figure 9 illustrates how the hybrid method embodies features of the baseline and scaled methods.

A repeated leave-one-out cross validation experiment is conducted to calculate the cross-entropy for each primary expression. During each iteration, 8 of 9 annotator intensity ratings are transformed using each of the 4 methods. The ratings are then used to estimate the parameters of a Beta distribution. The cross entropy H_{beta} is then calculated between the distribution and the left out intensity rating. A sample image with its respective entropy values for each method per expression is shown in Figure 12. The distribution of entropy values across all images and primary expressions using the 4 methods are shown in Figure 13. Exploratory results indicate that the weighted method best minimizes cross-entropy with new ratings and thus can be used to transform raw intensity ratings into tangible quantities for Beta fitting. Future work will consist of further determining the impact of each preprocessing method on the number of annotators needed and the performance of the benchmarking metrics. These insights will allow for improved guidance when utilizing preprocessing methods.

Appendix C. Expression Dimensionality Study

Our analysis of the compound expression hypothesis indicates that a 6-dim model for primary expressions can best model compound expression (Section 4.4 in the main paper). We further explored (1) how many dimensions, chosen from any of the 21 expression dimensions, are sufficient and (2) whether Ekman's 6 primary dimensions are sufficient. To explore these questions, we conduct an experiment using $N = 1000$ images where $L = 9$ annotators are instructed to provide ratings for $D = 21$ expression dimensions (see Fig. 3). The 21 dimensions are comprised of the 6 primary and 15 compound expressions.

To answer how many dimensions are necessary, we consider which subset of $k \in \{1, 2, ..., D-1\}$ expression(s) is best suited for predicting the remaining $D - k$ expression dimensions. We explore this through (a) a combinatorial approach (all-emotion combinatorial), where each possible combination of expressions is considered, and (b) a greedy approach, where the next best expression feature is iteratively added to the model. In this approach, the first expression in the greedy feature set is the single expression that best minimizes MAE when predicting the other 20 expressions. The next expression is then selected under the criteria that the resulting model best minimizes MAE when the expression is added to the greedy feature set over other models using an expression not in the greedy set. To explore whether the 6 primary dimensions are sufficient, a combinatorial approach limited to the 6 primary expressions is also considered, which we call the (c) primary combinatorial approach.

k.	Best Feature Combination	MAE %
1	disappointed	6.5
2	happy, betrayed	5.6
3	happy, spooked, disappointed	5.2
4	happy, sad, surprised, outraged	4.9
5	happy, sad, angry, surprised, contemptuous	4.6
6	happy, sad, angry, surprised, contemptuous, desperate	4.5
7	happy, sad, angry, surprised, contemptuous, desperate, hopeful	4.3
8	happy, sad, angry, surpr., contempt., hopeful, fearful, remorse.	4.2
...	...	...
15	happy, sad, angry, surpr., contempt., hopeful, fearful, disgusted, outraged, amazed, desperate, spooked, disbel., disapp., remorse.	3.6
...	...	...
20	D - cruel	3.3

Table 1. **Simplified representations**. We explore which subset of k expression dimensions are best suited as features to predict $D - k$ expressions. We consider all combinations of expressions and record the subsets that minimize the mean absolute error when a WLS Linear Regression model is fitted to predict the remaining expressions (all-emotion combinatorial approach, see Sec. C).

For each approach, we conduct a repeated 3-Fold weighted least squares (WLS) multi-output regression experiment to predict the mean, μ, of the Beta distribution for the remaining expression dimensions. Images are weighted by their average entropy across all expression dimensions S_i such that $w_i(S_i) = \exp(-S_i^2)$.

The best observed MAE using k expressions to predict $D - k$ expressions is illustrated in Figure 14 for each of the 3 approaches. To analyze the performance across the number of features used, we estimate a linear fit for each of the three methods. Using features 1 to 6, the slope of the fitted line is -0.003, -0.005, and -0.005 for the all-emotion combinatorial, primary combinatorial, and greedy methods respectively. Using features 6 to 21, the slope of the fitted line is -0.001 and -0.001 for the combinatorial primary and greedy methods respectively. It can be observed that very little is gained after using more than 6 expression dimensions.

Table 1 shows which specific expressions were best at predicting the remaining $D - k$ dimensions from the all-emotion combinatorial exploration. We express MAE as percent of the $[0, 1]$ dynamic range. Limiting to the primary expressions performs about as well as the best combination of any of the 21 expressions. Thus, our preliminary findings support that the primary 6 expression dimensions are sufficient.

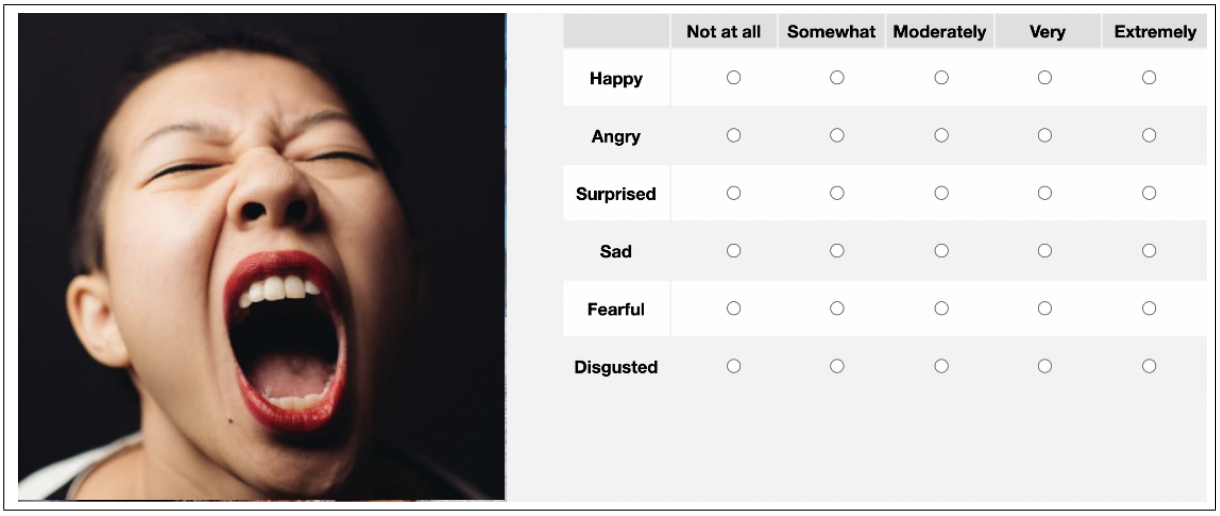


Figure 1. **Primary expression annotation collection.** A custom annotator interface is administered through Amazon SageMaker GroundTruth for collecting multi-dimensional graded expression annotations. SageMaker GroundTruth allows for efficient collection of annotations from both private workforces and Amazon Mechanical Turk workers. A response is required for each emotion category.

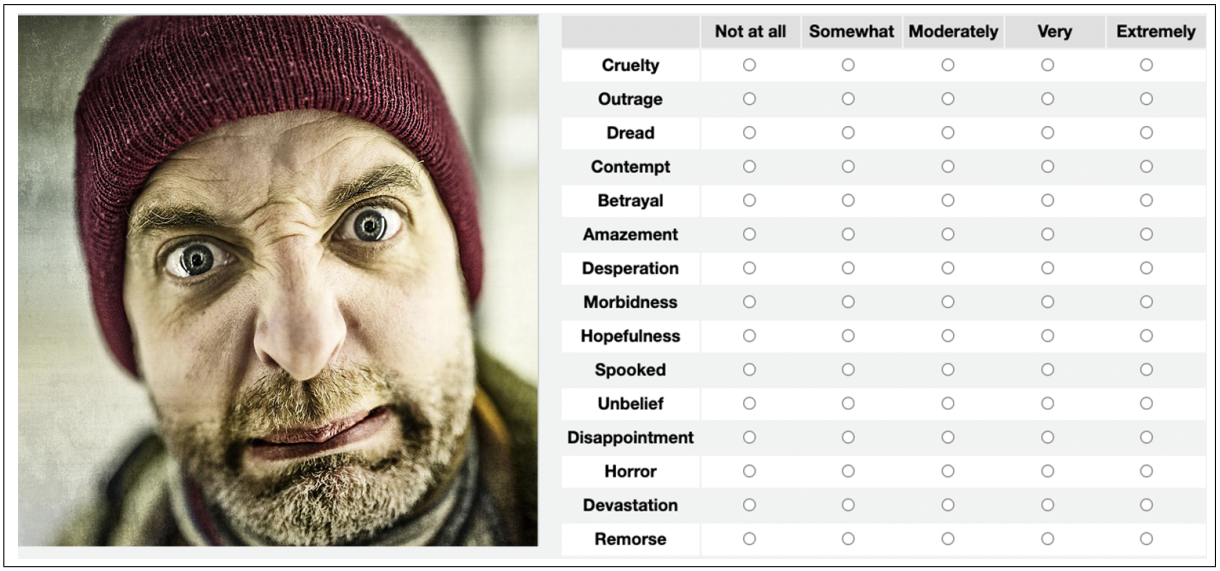


Figure 2. **Compound expression annotation collection.**A custom annotator interface for collecting multi-dimensional compound expression ratings along 15 dimensions.

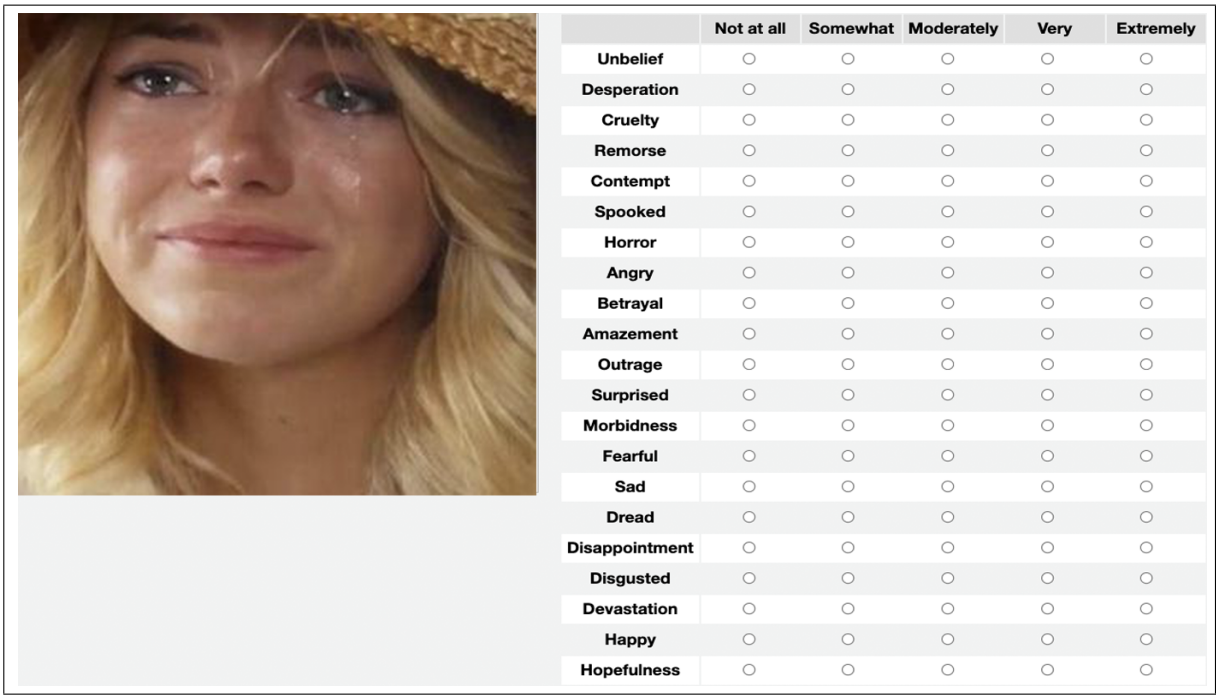


Figure 3. **21-dimension expression annotation collection.** A custom annotator interface for collecting multi-dimensional expression ratings along 21 expression dimensions. The order of expressions presented was randomly determined.

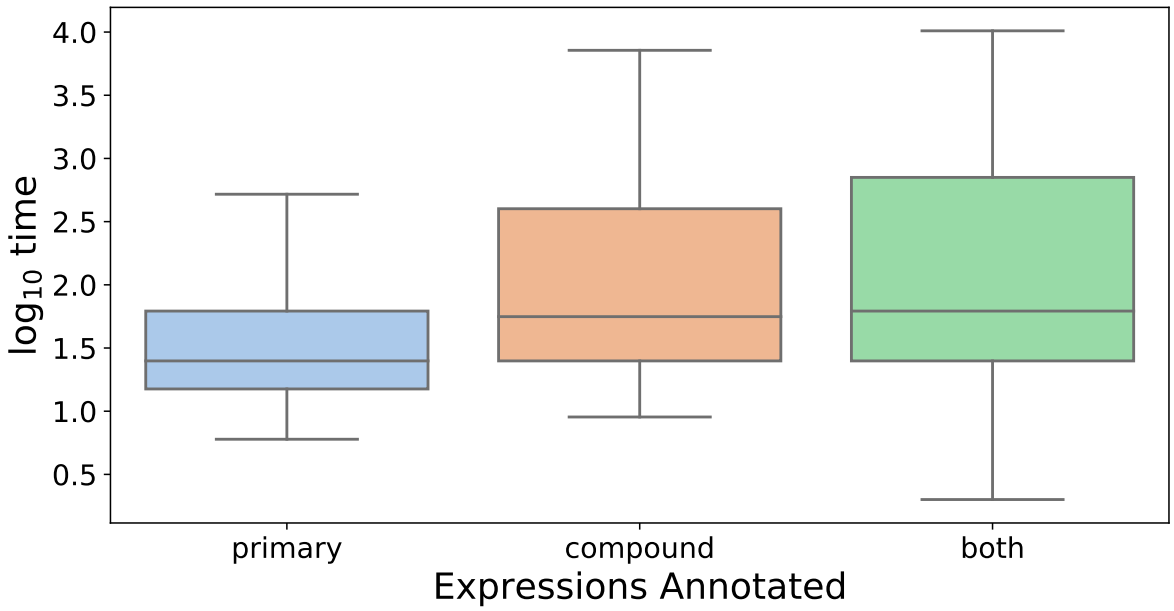


Figure 4. **Annotation Times.** The box plots measure annotation times per image for the 6 primary, 15 compound, and set of 21 expressions (as described in Section 4.1 and 4.2 of main paper). Time is displayed on a log_{10} scale.

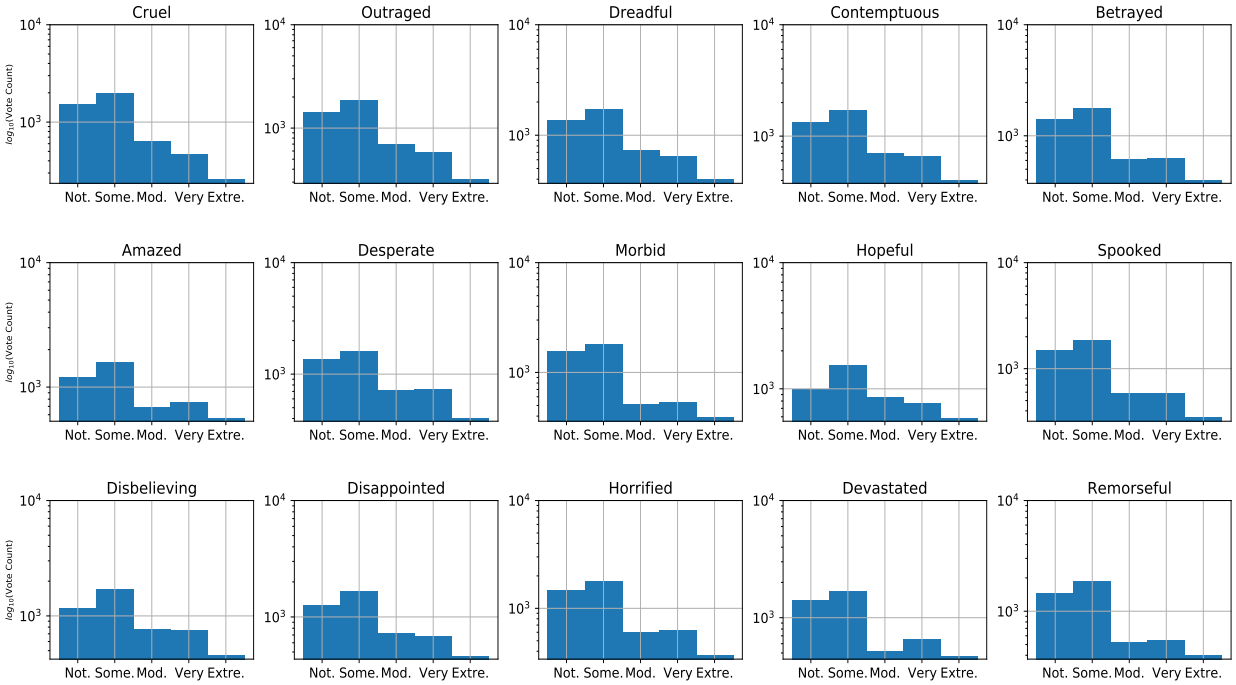


Figure 5. **Annotator rating frequency: Compound Expressions.** The distribution of annotations across all images per expression dimension during the 15 compound expressions experiment. (See Sec. 4.2)

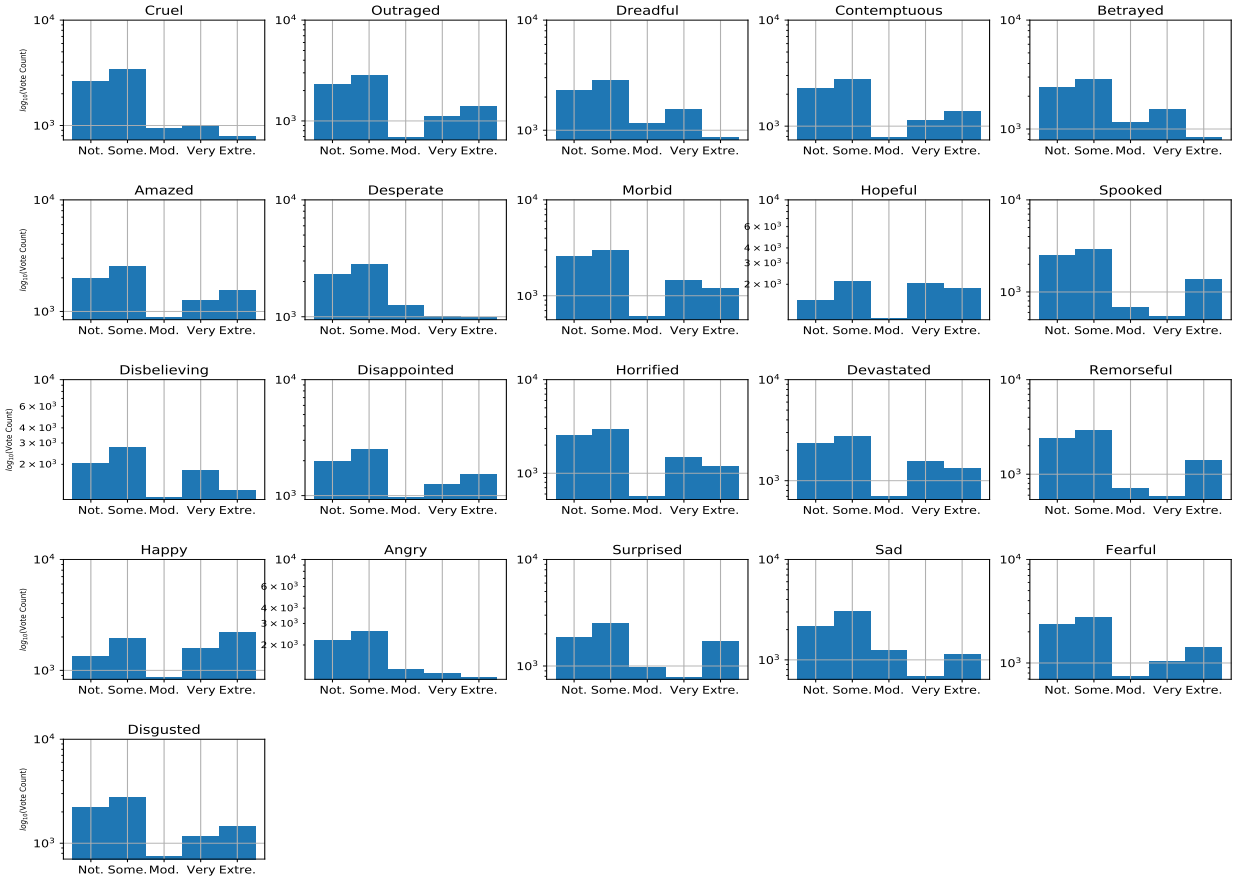


Figure 6. **Annotator rating frequency: 21 Expressions.** The distribution of annotations across all images per expression dimension during the 21 primary and compound expressions experiment. (See Sec. 4.2)

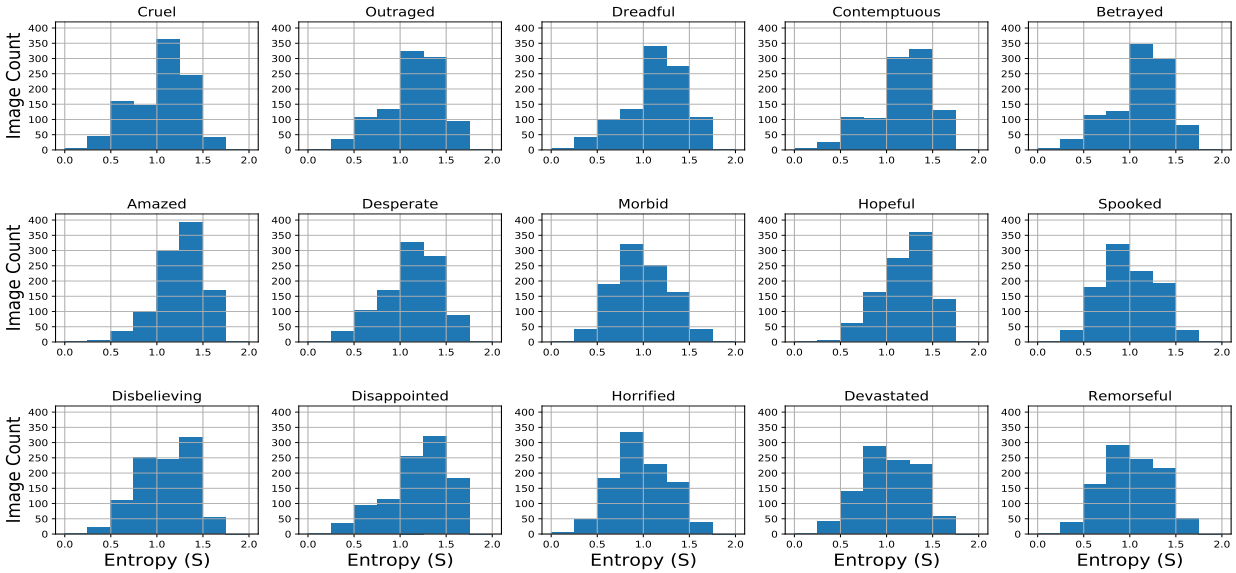


Figure 7. **Annotator agreement: Compound Expressions.** The frequency of entropy values (S as defined in Eq. 1, Sec. 3.3 of main paper) across all images and expressions annotated during the compound expression experiment. Lower entropy scores signify greater agreement between annotators. (Sec.4.2)

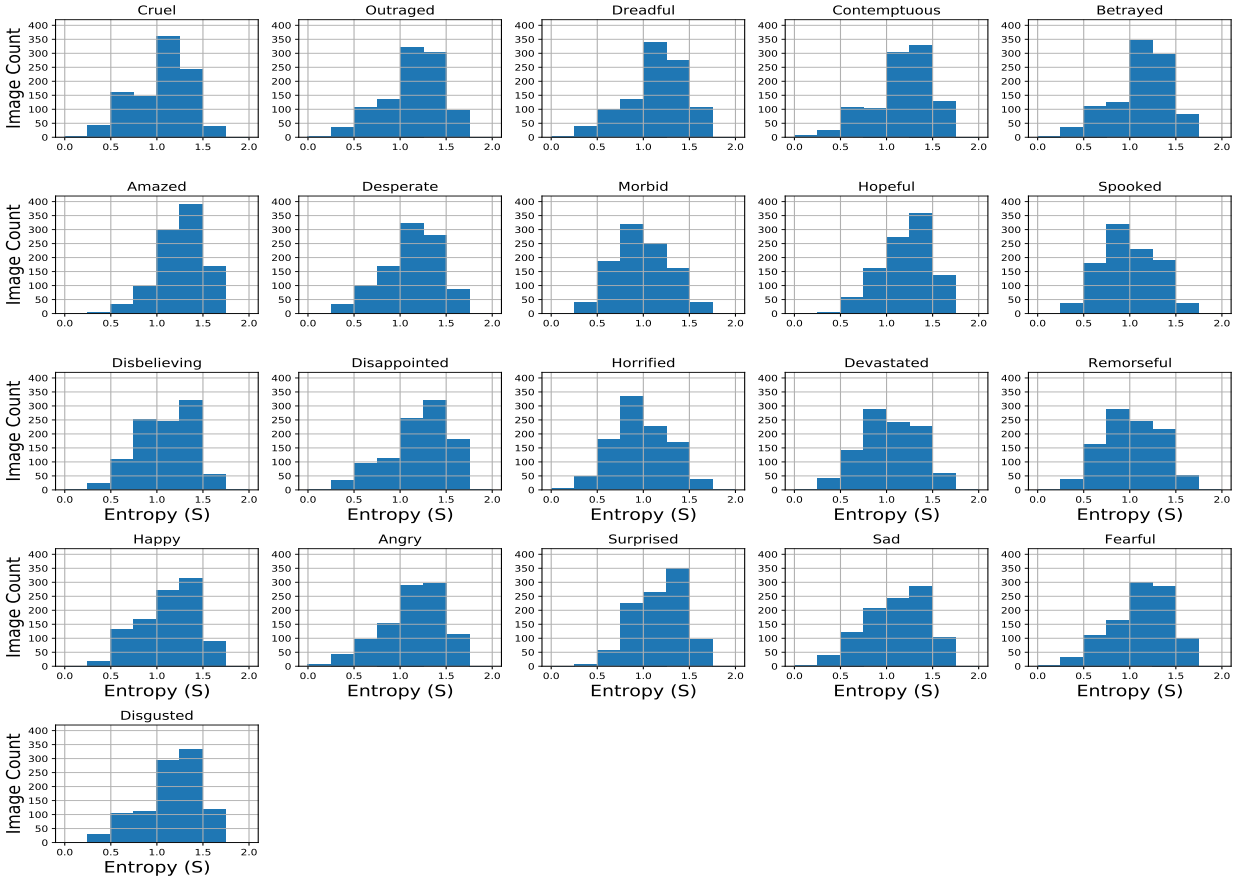
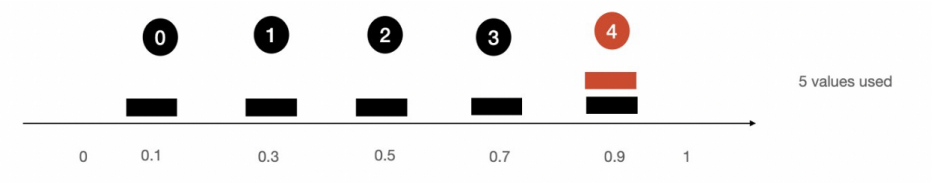


Figure 8. **Annotator agreement: 21 Expressions.** The frequency of entropy values (S as defined in Eq. 1, Sec. 3.3 of main paper) across all images and expressions annotated during the 21 primary and compound expression experiment. Lower entropy scores signify greater agreement between annotators. (Sec.4.2)

Normalized Method: Each raw intensity rating is normalized such that $r_{(i,d,l)} = (v_{(i,d,l)} / 5) + 0.1$

Baseline Method: Each bin is initialized with a single rating. Each annotator rating is then transformed using the normalization method.



Weighted Method: Each individual annotator rating is transformed into 4 ratings such that $W(r_{(i,d,l)}) = [(r_{(i,d,l)} - 0.1),\ r_{(i,d,l)},\ r_{(i,d,l)},\ (r_{(i,d,l)} + 0.1)]$

Hybrid Method: Each bin is initialized with a single vote. Each vote is then transformed using the weighted method.

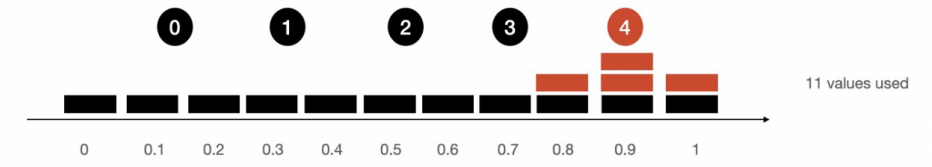


Figure 9. **Four data conditioning methods.** Four methods to condition raw annotator ratings prior to estimating the parameters of a beta distribution are considered. The normalization method was used during all analyses conducted in the main paper (Sec. B).

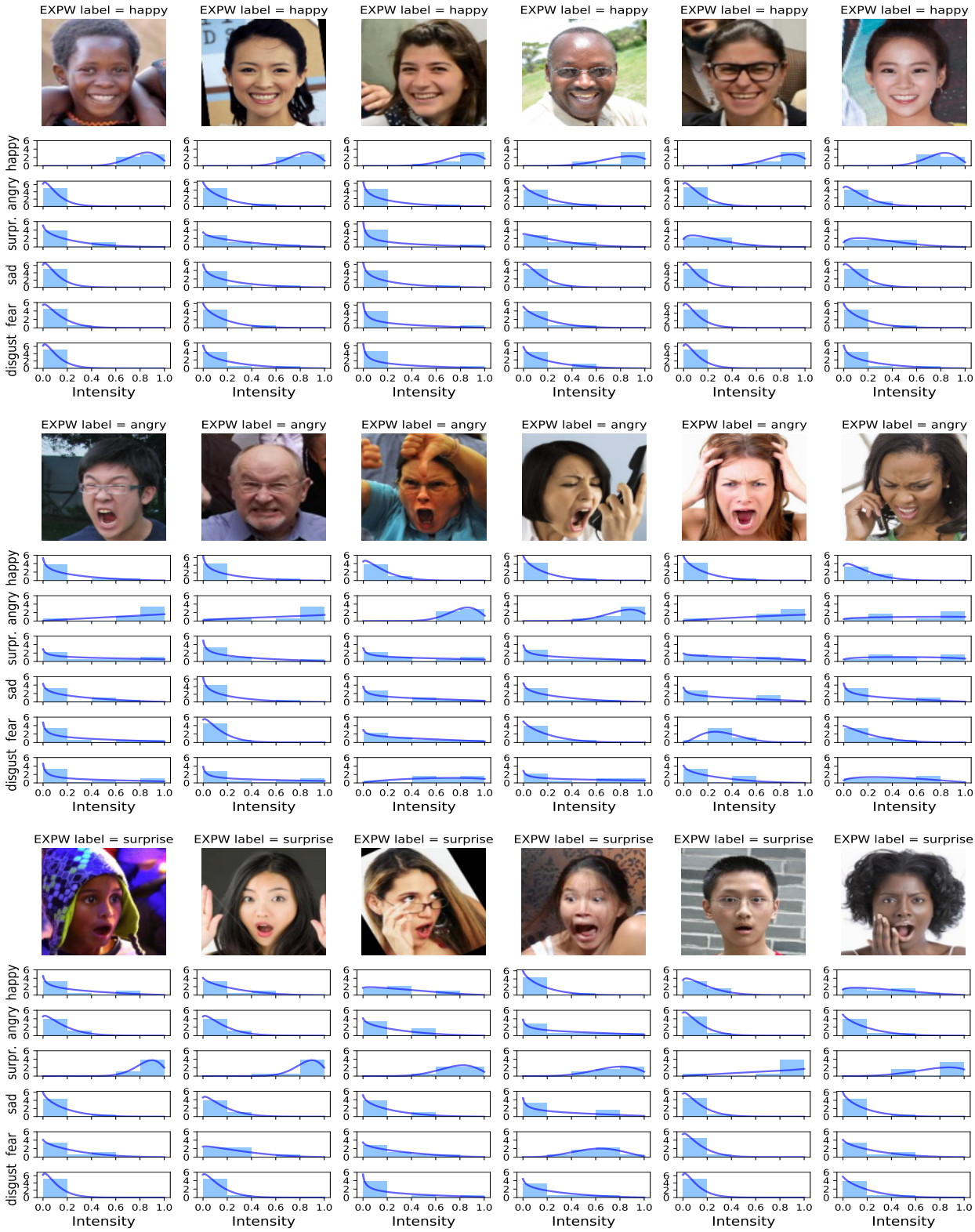


Figure 10. **EXPW sample images: happy, angry, and surprise.** Sample images obtained from the ExpW dataset that comprise the 1000 image set. The original EXPW expression label is shown above each image. The histogram of annotator ratings is shown for each expression category along with the Beta distribution obtained from fitting the ratings.

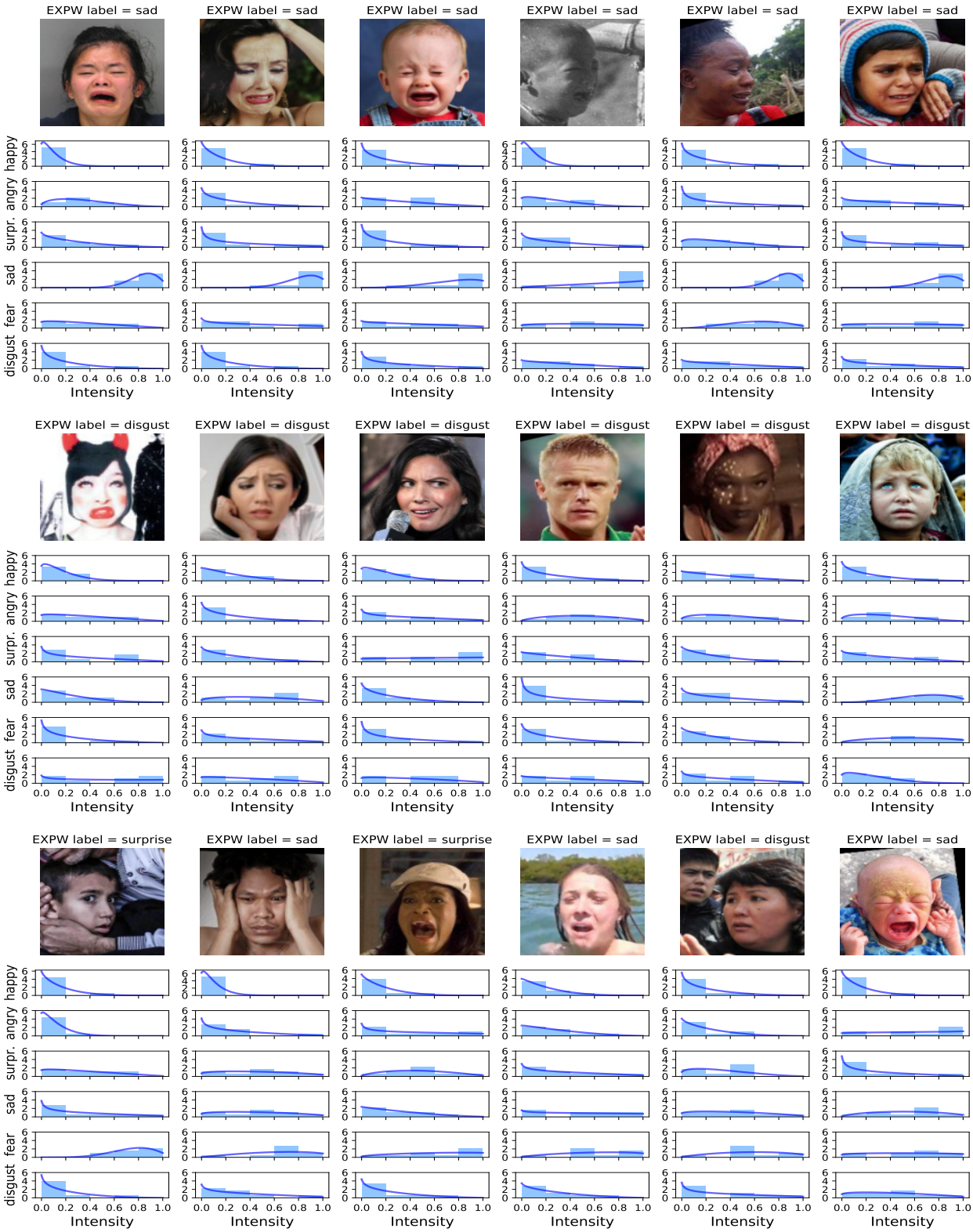


Figure 11. **EXPW sample images: sad, disgust, and fear** Sample images obtained from the ExpW dataset that comprise the 1000 image set. The original EXPW expression label is shown above each image. The histogram of annotator ratings is shown for each expression category along with the Beta distribution obtained from fitting the ratings. For the 'fear' expression, images are shown that had high annotator ratings for fear and another expression label in EXPW.

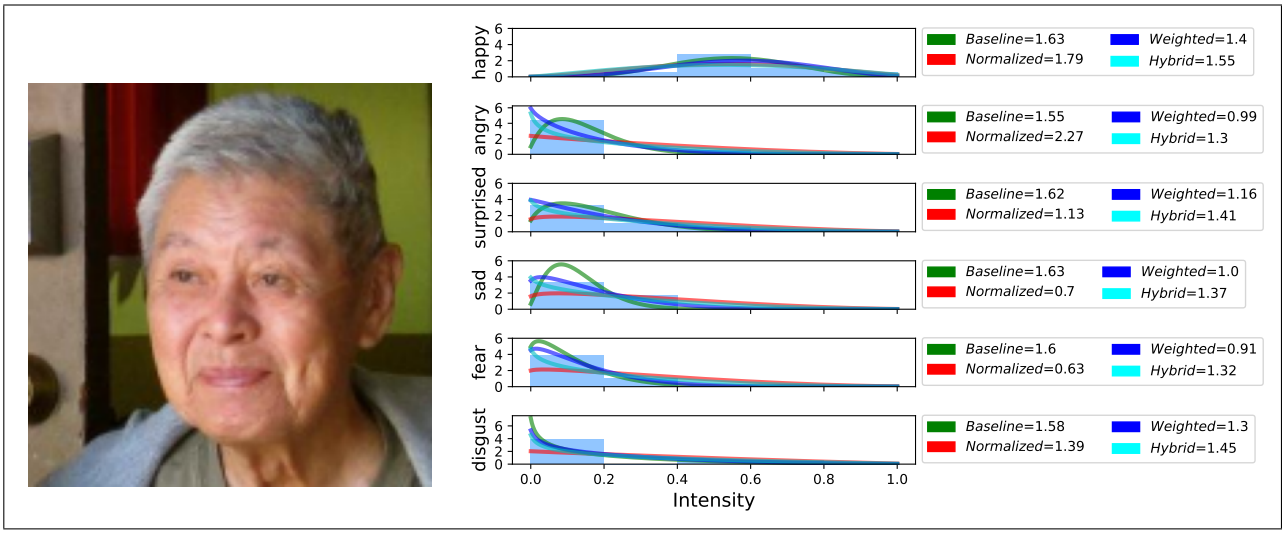


Figure 12. **Comparing Beta fitting methods.** We consider the best of 4 methods to condition raw annotator intensity ratings through a repeated leave-one-out cross-entropy validation experiment (see Sec. B). The cross-entropy score, H_d, for each transformation method per expression d is shown. Lower cross-entropy scores suggest the distribution fits new annotator ratings well.

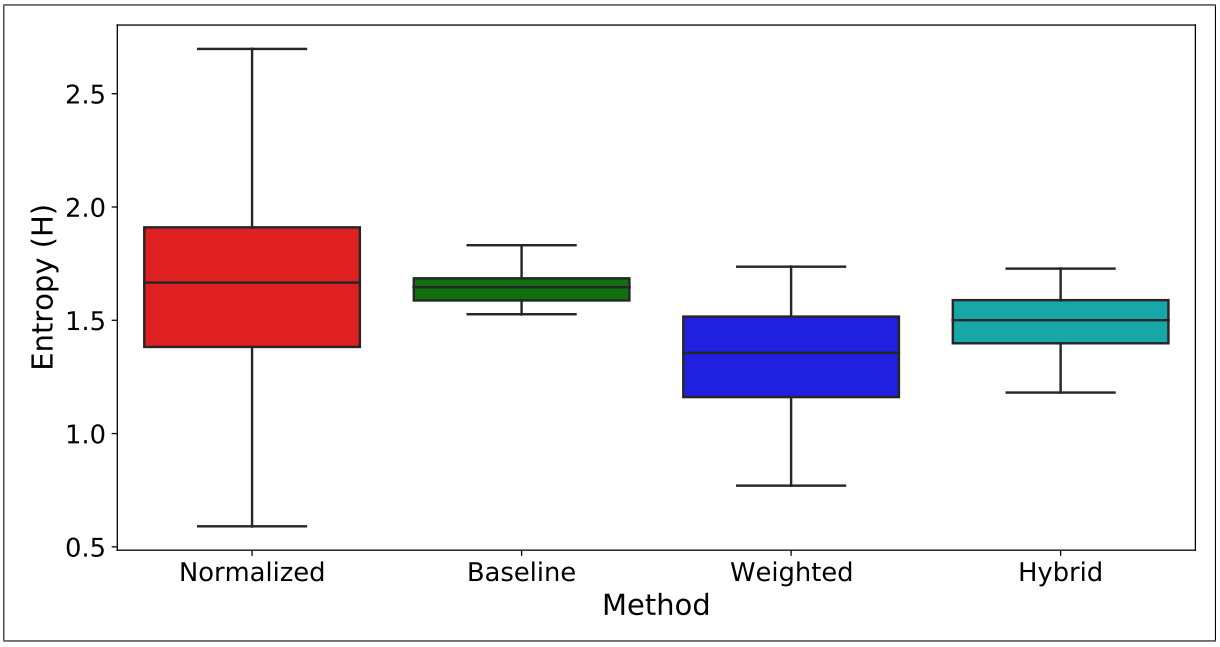


Figure 13. **Intensity preprocessing methods.** We compare 4 methods to transform raw annotator ratings to estimate the parameters of a Beta distribution (Sec. B). The cross-entropy values are calculated using a repeated leave-one-out validation experiment of annotator intensity ratings. The box-plot shows the cross-entropy values across all expressions and images. The weighted method best minimizes the cross-entropy between a distribution and new annotator ratings.

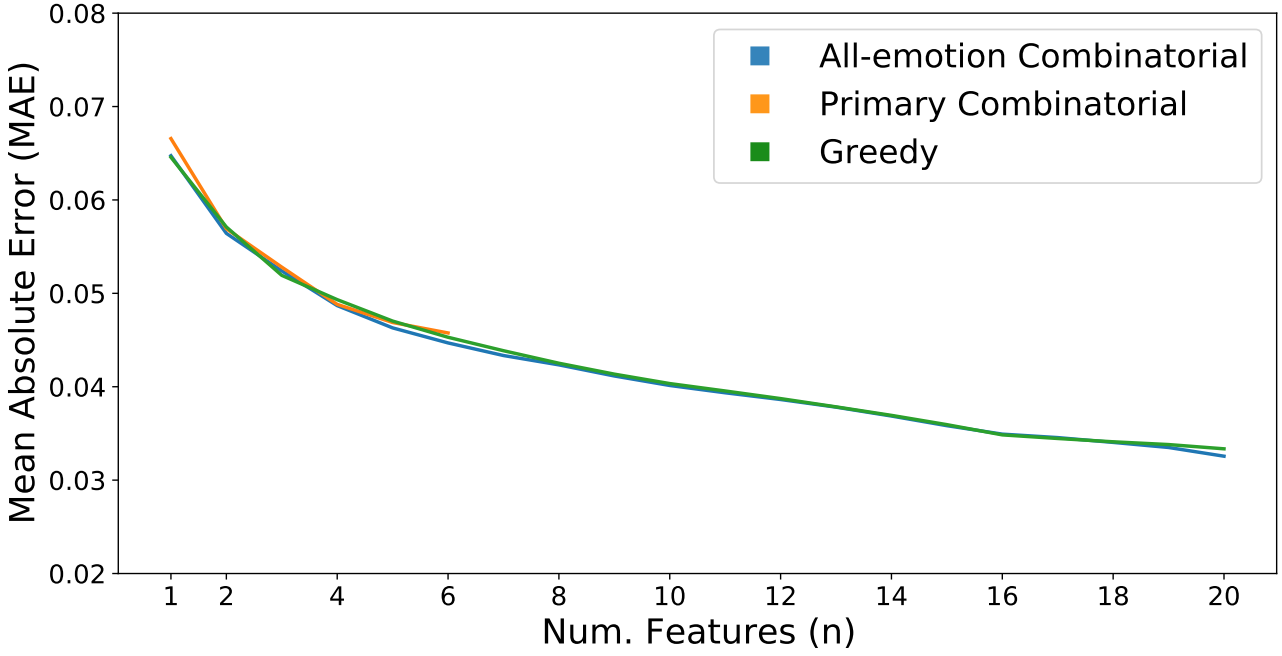


Figure 14. **Dimensionality Analysis.** Three approaches were used to examine the behavior of models that use k expressions to predict $D-k$ expressions (Sec. C). The MAE for each model using k expressions per method is considered. We find that 6 dimensions is sufficient. The 6 primary emotions perform as well as the model using the 6 best expression features.