Multi-Dimensional, Nuanced and Subjective – Measuring the Perception of Facial Expressions

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

Abstract

Humans can perceive multiple expressions, each one with varying intensity, in the picture of a face. We propose a methodology for collecting and modeling multidimensional modulated expression annotations from human annotators. Our data reveals that the perception of some expressions can be quite different across observers; thus, our model is designed to represent ambiguity alongside intensity. An empirical exploration of how many dimensions are necessary to capture the perception of facial expression suggests six principal expression dimensions are sufficient. Using our method, we collected multidimensional modulated expression annotations for 1,000 images culled from the popular ExpW in-the-wild dataset. As a proof of principle of our improved measurement technique, we used these annotations to benchmark four public domain algorithms for automated facial expression prediction.

1. Introduction

Humans communicate using their body. Automating the perception of bodily and vocal expressions is necessary towards building machines that can interact gracefully with humans [8, 59]. Facial expression is an important channel of the communication [9, 16], and perception of facial expressions is important for social interaction [16, 29]. Computer vision researchers have long been interested in measuring human facial expressions from images and video [4, 12, 41, 47, 56] with the aim of replicating it in machines [4, 41].

Automated facial analysis is rooted in machine learning, thus model training and benchmarking rely on large well-annotated datasets. This raises three questions which are not well addressed in the facial expression perception literature: First, how should images be annotated, i.e. what is a good representation of human perception of facial expression? Second, can we measure the perception reproducibly

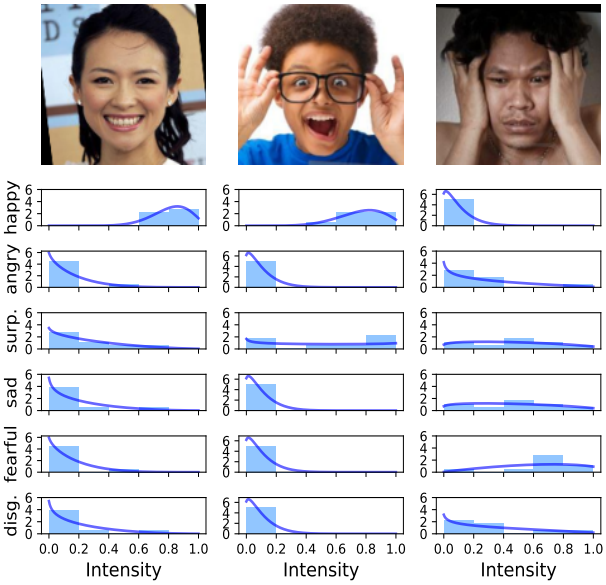

Figure 1. **Face expression perception is multidimensional, nuanced and subjective.** Perceived expression is measured by asking 9 crowdsourced annotators to report perceived intensity for each of six dimensions (15 and 21 dimensions in other experiments as described in Sec. 4.1). Annotation histograms and Beta distribution fits (Sec. 3.2) are shown. One dimension, 'happy', captures the expression of the first face. The second requires two: 'happy' and 'surprised'. The third requires more dimensions and is more ambiguous.

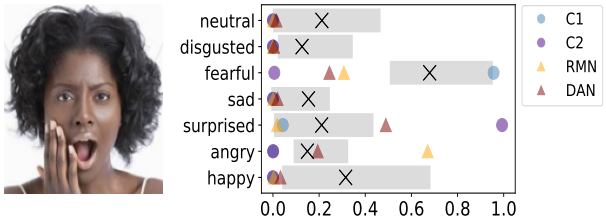

Figure 2. **Expression predictions compared to ground truth.** The outputs of 4 expression prediction algorithms (colored symbols) are compared to the *ground truth* obtained from our annotations for one image. Gray bands: confidence intervals of our ground truth, ×: μ (See Sec. 3.2 and 4.3).

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

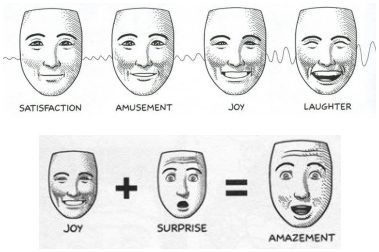

Figure 3. **Evidence for intensity and compound expressions** found in the arts. (Top) Expressions vary in intensity. (Bottom) Any pair of primary expressions may be combined to obtain a valid composite expression (Table 1, Sec. 2.1). (Adapted with permission from Scott McCloud's *Making Comics* [42], pages 84-85).

and, yet, efficiently? Third, what is the right metric to compare an algorithm's output with human ground truth?

To answer the first question we rely on insight from experimental psychology [15, 25, 31, 45], which suggests that facial expressions are multidimensional and are perceived with different degrees of intensity (see Fig. 1). This constrasts with the common practice in computer vision where expressions are often annotated as a binary and/or a one-hot code (Sec. 2). To answer the second question, we focus on crowdsourcing – we annotated 1,000 ethnically diverse in-the-wild faces from a public dataset [60], where each was rated by 9 annotators for each of the 6 primary Ekman expressions [17], the corresponding 15 compound expressions [42, 49] and the combined set of 21 expressions. On this dataset we explore the consistency of annotations, as well as the number of dimensions that need to be annotated. Lastly, we propose a metric that compares human annotations, including their ambiguity, with algorithm prediction, and use our annotated dataset to benchmark four recent algorithms (Sec. 4.3).

Our main contributions are: **1.** An efficient method for collecting and modeling reproducible, multi-dimensional, modulated facial expression annotations crowdsourced from humans. The novel modeling technique transforms annotations into probability distributions to express measures of expression intensity and ambiguity. **2.** A benchmark for facial expression prediction algorithms consisting of annotations on 1,000 face images from a public in-the-wild dataset, and a metric for algorithmic accuracy of expression prediction.

2. Related work

2.1. Expression and perception

The *mechanics and repertoire of facial expression* are fairly well understood. Ekman [15] postulated six primary dimensions (happy, angry, surprised, sad, fearful, disgusted), independent of culture and experience, and described the muscle actions that produce such primary expressions [17]. This point of view has, by and large, stood the test of time [31]. Artists have empirically observed that some expressions involve simultaneously more than one of the six primary dimensions [42]. Recent research has explored that intuition and characterized the corresponding facial actions [13].

Expression 1	Expression 2	Compound Hypothesis
Angry	Happy	Cruel
Angry	Surprised	Outraged (P)
Angry	Fearful	Dreadful*
Angry	Disgusted	Contemptuous (P), Outraged (M)
Angry	Sad	Betrayed
Happy	Surprised	Amazed
Happy	Fearful	Desperate
Happy	Disgusted	Morbid (P)
Happy	Sad	Hopeful
Surprised	Fearful	Spooked
Surprised	Disgusted	Disbelieving (P)
Surprised	Sad	Disappointed
Fearful	Disgusted	Horrified
Fearful	Sad	Devastated
Disgusted	Sad	Remorseful

Table 1. **The compound expressions hypothesis.** All pairwise combinations of 6 primary expressions are considered. The names of the compound expressions were obtained from Scott McCloud's *Making Comics* [42] and Robert Plutchik's theory of expression [49]. McCloud nor Plutchik use multiple words to describe the compound of angry and fearful. We use 'Dreadful' here following prior work [33, 38]. An analysis of whether complex expressions may be suitably modeled as superposition of primary expressions can be found in Sec. 4.4. See also Sec. 2.1 and Fig. 4.

The *perception of facial expressions* is also well studied, alongside the perception of other socially relevant attributes such as gender, age, and trustworthiness. Human annotators can make fast judgments [54], which may be used in important decisions [50]. Often, there is good agreement amongst annotators on their perception, although some annotators are believed to be more perceptive and reliable [25]. Whether and when the annotators' judgments correspond to meaningful and useful information is still debated [30]. Notably, Barrett et al. [2] questioned whether a person's internal emotional state may be inferred from facial expression, an assumption central to Ekman's earlier views on emotion and facial expression. Our method and data will help provide empirical evidence for the ongoing Barrett-Ekman debate.

2.2. Automating prediction of expression perception

Computer vision algorithms may be used to *measure facial expressions and predict the voluntary report that a regular person would make of their perception of the expression* upon looking at a face image. We use the shorthand "prediction of expression" and "expression classification", and the meaning should be clear by context.

Detecting and analyzing human faces was recognized as an important task from the inception of computer vision [28]. Since the early 1990s, computer vision researchers have been interested in automating the perception of facial expressions [4, 12, 41, 47, 56] and deep learning is

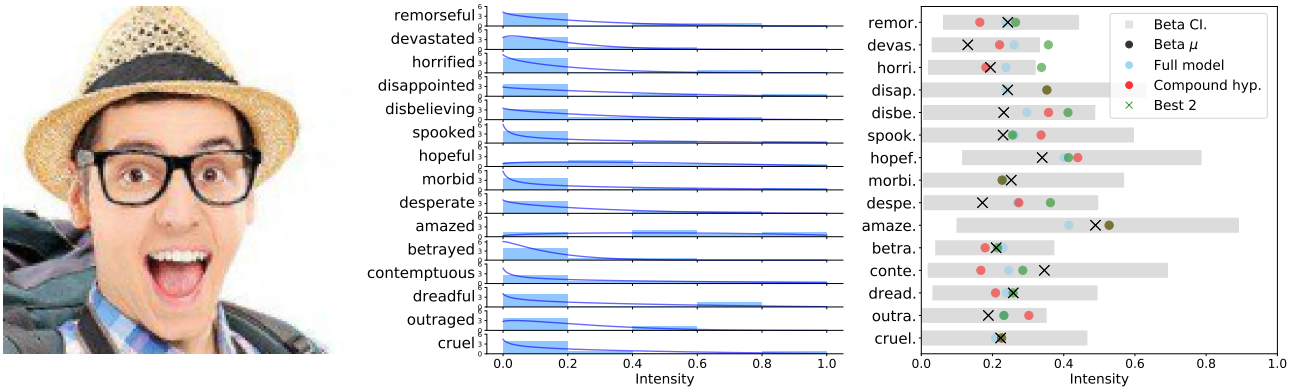

Figure 4. **Predicting compound expressions from the primary six.** We trained three regressors to predict compound expressions from those of primary ones. The first regressor (full model) utilizes all 6 primary expressions as features to predict each compound expression. The second (compound hyp.) uses 2 primary expressions to predict each compound expression, where the two expressions are taken from the compound expressions hypothesis (Table 1). The third model (best 2) uses the two primary expressions that best fit each compound expression. The μ of the fitted beta distributions are also plotted (see Sec. 3.2). The gray confidence intervals are computed from the beta distribution and represent 66% confidence. Goodness-of-fit measurements comparing the different regressors are discussed in Sec. 4.4.

the prevalent approach of modern algorithms [1, 19, 34, 52, 55]. We make no direct contribution to algorithms. Rather, we focus on defining a minimal and sufficient representation for expression perception, methods for dataset annotation and methods for benchmarking algorithms.

2.3. Data annotation of perceived expression

Datasets for training and testing automatic models are annotated by human observers who classify the expressions they perceive. There are two popular approaches. The first measures *facial actions* [4, 17], i.e. the contraction of specific muscles, a demanding task even for experts. Once the facial actions are measured, one may infer the underlying expression. The second, which we adopt, aims at making use of the perception of regular people and thus focuses directly at naming facial expressions. The perception of many face attributes is empirically consistent across human annotators [24, 31], which justifies considering it as an intrinsic property of the image. However, a number of questions are left open in the literature.

First, crowd-sourcing, e.g. on Amazon Mechanical Turk (AMT), has been adapted in the domain, but the consistency of reports in such setting, has never been studied to the best of our knowledge. Goodfellow [21] discussed observations on human performance for expression classification but not for crowdsourced settings and not through formal experiments. Recent work [35] compared annotations from different annotators with the goal of consolidating scores, yet did not quantify the reliability of the annotation mechanism. Our work directly analyzes the reproducibility of crowdsourced annotations, and provides practical recommendations on how many expression dimensions and how many annotators are needed.

Second, it is not known whether all images are equally consistently interpreted, and whether some are ambiguous.

Dataset	N Annotated	In the wild	Multi-Dim. expressions*	Modulated raw annotations	Modulated aggregated annotations
AffectNet [44]	420K	✓	7	✗	✗
EmotioNet [18]	50K	✓	6+17‡	✗	✗
FERPlus [3]	36K	✓	7	✗	✗
iCV-MEFED [22]	31K	✗	✗	✗	✗
RAF-DB [36]	30K	✗	7+11	✗	✗
ExpW [60]	8.3k	✓+	✗	✗	✗
RAF-ML [35]	4.9k	✓	6	✗†	✓††
RADIATE [7]	1.7k	✗	✗	✗	✗
CAFE [37]	1.2k	✗	✗	✗	✗
Our study	1k	✓	6+15	✓	✓
NimStim [51]	0.7k	✗	✗	✗	✗
BU-3DFE [58]	0.6k	✗	✗	✓	✗
Dawel, Amy, et al [10]	0.6k	✗	✗	✓	✓
NIMH-ChEFS [14]	0.5k	✗	✗	✓	✗
Karolinska [20]	0.5k	✗	✗	✓	✗
DEFSS [43]	404	✗	✗	✓	✗
RaFD [32]	216	✗	✗	✓	✗
JAFFE [39]	213	✗	7	✓	✗
Chicago [40]	158	✗	✗	✓	✗

Table 2. **Synopsis of face image collections annotated for expression.** (*) Primary and compound expressions. (‡) EmotioNet annotated facial action units to generate expression categories. (+) ExpW faces are sourced from both movies and in-the-wild images. (†) RAF-ML asked each annotator to select one expression out of all categories, the rating is binary per category. (††) RAF-ML obtained modulated labels by combining binary annotations from a large number of annotators. (Sec. 2.3)

We address this question and find that, indeed, some face images are ambiguous, which motivates representing the ground truth as an interval, or a probability distribution, rather than a single label or value (Sec. 4.2).

Third, while it is well established in the psychology literature that facial expressions differ in intensity and multiple primary expressions (or compound expressions) can coexist [13, 42, 49], none of the existing datasets that one may use for testing facial expression algorithms support both *multi-class encoding* and *modulated intensity of each expression class* (see Table 2). We propose here a *method for collecting annotations of multidimensional modulated expressions for each face image.* The method used to generate the RAF-ML dataset [35] produced multi-dimensional modulated annotations and is most similar to ours; however, they asked annotators to pick only one expression per face, while we allow annotators to pick an intensity rating for each expression per face. Furthermore, their method required 40 annotators per face while in our case we find that 5-6 annotators are needed (Sec 4.2), thus, our method is less expensive and more likely to capture subtle expression variations. Blank et al. [5] also introduced a novel method to estimate multi-dimensional intensity ratings from discrete one-hot annotations using co-occurrence matrices. However, it has not yet been shown that such approaches are an equivalent or sufficient representation of how a person would perceive multiple expressions in a face.

Fourth, while it has been suggested that compound expressions may be thought of as combinations of the six primary expressions [13, 42] (see Table 1), it has not yet been verified whether the perception of crowdsourced annotators satisfies this hypothesis, and thus whether one may simply annotate the six primary expressions, or whether both primary and compound (15 pairwise compounded expressions) need to be annotated (for a total of 21 expressions). This question is explored in Sec. 4.4.

2.4. Benchmarking perceived expression

Most previous methods for benchmarking facial expression algorithms are based on comparing the output of the algorithm to a one-hot encoding [1, 6, 19, 52, 55]. Recently, RAF-ML [35] proposed the direct usage of several common metrics for evaluating multi-dimensional measurements onto expression prediction. Along this direction, we further propose a new metric where not only modulated algorithmic predictions are compared with modulated ground truth annotations via a distance metric, but also the ground truth ambiguity is represented as a probability distribution and the algorithm's predictions are matched to such distribution using a cross-entropy metric.

3. Methods

3.1. Multidimensional modulated annotation

People may perceive one or more primary expressions at varying intensities when viewing a face. To adequately capture this phenomenon, we designed three graphic interfaces where the annotator may report how much or how little they perceive each facial expression on a 5-point scale (Fig. 5). Throughout our annotation collection, we adapted three sets of multi-dimensional expression options in our interface: (1) *6 primary expressions* ('happy','angry','surprised','sad','fearful','disgusted'), (2) *15 compound expressions* ('cruel', 'outraged', 'dreadful', 'contemptuous', 'betrayed', 'amazed', 'desperate', 'morbid', 'hopeful', 'spooked', 'disbelieving', 'disappointed', 'horrified', 'devastated', and 'remorseful'), and (3) the set of 21 primary and compound expressions.

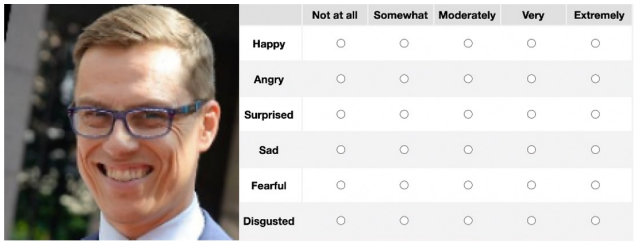

Figure 5. **Annotator interface for expression annotation.** The interface shown here was used to annotate the perception of the expression of the six primary expression classes (see Sec. 3.1 and 4.1). Annotators must select one intensity (column) per dimension (row). A similar interface was used for annotating the 15 compound expressions (Table 1) and the set of 21 primary and compound expressions separately (see Supplementary).

We took an online crowdsourcing approach to collect the human annotations because it is easily accessible, scalable, and cost-efficient. The custom interface was developed using Amazon SageMaker Ground Truth, which allows one to easily crowdsource the task.

The experimental details on curating the dataset are described in Sec. 4.1. To measure the reliability and efficiency of the online crowdsourcing approach, we then analyze the annotator efficiency and rating consistency in Sec. 4.2. A benchmark analysis on four public-domain algorithms is presented in Sec. 4.3, and an exploration of modeling compound expressions is introduced in Sec. 4.4.

3.2. Modeling annotations with Beta distributions

After multidimensional modulated annotations have been collected, they can be used to model the variability in human perception of a given expression. From the annotation interface (Fig. 5), we map the modulated intensity choices "Not at all", "Somewhat", ..., "Extremely" to numerical values $0.1, 0.3, ..., 0.9$. We denote the modulated annotator intensity rating with $r_{(i,d,l)} \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, where $i \in [1, 2, ..., N]$ indicates the image index ($N = 1,000$), $d \in [1, 2, ..., D]$ is the expression dimension (D =6, 15 or 21), and $l \in [1, 2, ..., L]$ indicates the human label annotator index (typically $L = 9$).

In order to model expression perception intensity (magnitude) and spread (uncertainty) of the annotators' perceptions, we fit a Beta distribution [26] (Sec. 4.1) for each set of annotators' ratings $\mathbf{r}_{(i,d)} = \{r_{(i,d,1)}, r_{(i,d,2)}, ..., r_{(i,d,L)}\}$

for each expression dimension d of each face image i. We fit the ratings to the beta distribution $g(r|\alpha_{(i,d)}, \beta_{(i,d)})$ using maximum likelihood estimation (MLE) [46] obtaining parameters $\alpha_{(i,d)}$ and $\beta_{(i,d)}$. A small uniform regularization noise sampled from $[-0.1, 0.1]$ is added to each rating before computing the MLE. Omitting the subscripts for simplicity, the two parameters α and β control the shape of the curve, i.e. the position of the mode (or anti-mode) and the dispersion (or variance). The mean, μ, of the distribution is defined as $\mu = \frac{\alpha}{\alpha+\beta}$. Confidence intervals visualize the range of 68.3% (median $\pm\sigma$) of the distribution.

The ratings we collect from annotators are discrete values while the Beta distribution is continuous. Therefore, we approximate the definite integral of $g(r|\alpha, \beta)$ following the principle of the trapezoidal rule [57] with the resolution of the partition being the same as the five-level rating. This approximation yields a discrete version of the Beta function which we denote as $\bar{g}(r|\alpha, \beta)$ and use for calculation.

3.3. Annotation ambiguity and cross entropy

Annotators, based on their perception, may disagree on the intensity of each facial expression for the same face. We adapt a formulation of entropy to measure such intensity ambiguity amongst annotators. For each face image i per expression dimension d, we take a set of L ratings $\mathbf{r}_{(i,d)} = \{r_{(i,d,1)}, r_{(i,d,2)}, ..., r_{(i,d,L)}\}$, and calculate the entropy S as

$$S(\mathbf{r}_{(i,d)}) = \sum_{r=0.1}^{0.9} f(r, \mathbf{r}_{(i,d)}) * \log(f(r, \mathbf{r}_{(i,d)})) \quad (1)$$

where $f(r, \mathbf{r}_{(i,d)})$ indicates the frequency of rating value $r \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ in the set of L intensity ratings in $\mathbf{r}_{(i,d)}$. We apply the entropy to analyze the extent to which there is a consensus amongst the annotators and find the prevalent perception (Sec 4.2).

We introduce a cross entropy measure for comparison between a Beta distribution fitted from a set of ratings and another single intensity rating. If the latter comes from a human annotator, then we have $r \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$; if the latter comes from a model's prediction, we quantize it into the discrete values for fair cross-entropy calculation. To this point, we define the cross entropy between a discrete rating r and a Beta distribution given by α and β as:

$$H_\beta(r|\alpha, \beta) = -\log(\bar{g}(r|\alpha, \beta)), \quad (2)$$

which we use in analysis and benchmarking (Sec. 4).

3.4. Algorithm benchmarking metrics

Multidimensional modulated annotations can be used to benchmark current expression perception algorithms. The model predicted facial expression intensity for image i along expression dimension d can be denoted as $\hat{r}_{(i,d)} \in (0, 1)$. We calculate the cross-entropy H_d for an expression dimension d across all images using:

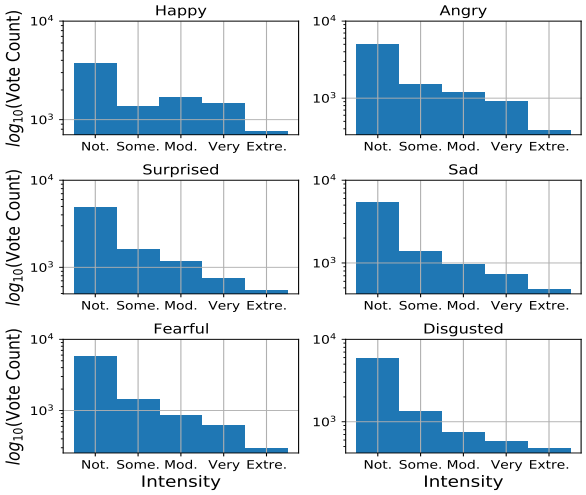

Figure 6. **Annotator rating frequency.** The distribution of annotations across all images per expression dimension (6 primary expressions). A given expression is most often not present on a given face.(Sec. 4.2)

(4).pdf

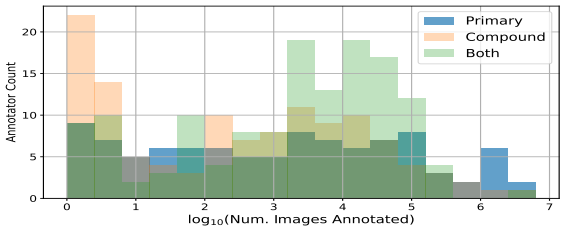

Figure 7. **Quantity of annotations by annotator.** The number of images rated by unique annotators. The x-axis is number of annotated images presented in a log_{10} scale. (Sec. 4.2)

$$H_d = \frac{1}{N} \sum_{i=1}^{N} H_\beta(\hat{r}_{(i,d)}|\alpha, \beta), \quad (3)$$

where $H_\beta(r|\alpha, \beta)$ is defined in Eq. 2. We also introduce M_d, a measure of absolute distance between the Beta distribution and the rating, for each expression dimension d as:

$$M_d = \frac{1}{N} \sum_{i=1}^{N} |\mu_{(i,d)} - \hat{r}_{(i,d)}|, \quad (4)$$

where $\mu_{(i,d)}$ is the distribution mean for image i along dimension d.

4. Experiments and analysis

The following subsections introduce the experimental design used to validate multidimensional modulated annotations (Sec. 4.1), assess annotator behavior during the task (Sec. 4.2), demonstrate their potential for use to benchmark expression perception algorithms (Sec. 4.3) and test the compound expression hypothesis (Sec. 4.4).

4.1. Experimental details

Image selection for benchmark annotation. 1,000 unique images were curated from the pre-existing Expression in-

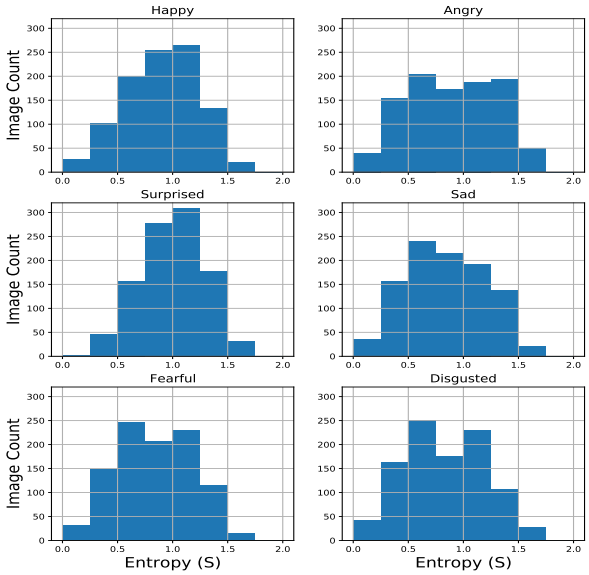

Figure 8. **Annotator agreement.** The frequency of entropy values (S as defined in Eq. 1, Sec. 3.3) across all images and expressions annotated during the primary expression experiment. Lower entropy scores signify greater agreement between annotators. (Sec. 4.2)

the-Wild (ExpW) dataset [60]. We first excluded images that contained non-human faces, a watermark corrupting the face area, and images with low resolution. To diversify the range of individuals in the set, we then filtered the data using ethnic and nationality keywords found in the original ExpW query metadata (i.e., African, American, Asian, European, etc.), leaving 42,790 images. We obtained face bounding boxes from the images and sampled faces by given ExpW expression labels to ensure a variety of facial expressions. Finally, 1,000 ethnically-diverse cropped face images above 40KB and of high image quality [11] were selected for inclusion.

Annotation collection. We designed a custom Amazon SageMaker GroundTruth graphic interface (Fig. 5) to collect annotator data in three experiments. First, participants were recruited to rate face images by the 6 primary expressions and paid at a rate of $0.072 per completed image. In the second experiment, participants rated images by the 15 compound expressions and were paid $0.24 per image. In the final experiment, participants rated images by the set of 21 expressions and were paid $0.24 per image. The median time taken to annotate 6, 15, and 21 expressions was 25, 56, and 62 seconds respectively, yielding an average pay rate of $13.23 per hour. Each image was annotated by 9 unique AMT participants per experiment, resulting in approximately 27,000 annotations and 358.3 hours of work.

Beta distribution fitting. Annotator intensity ratings were used to estimate shape parameters, α and β, of a Beta distribution per expression per face (Sec. 3.2).

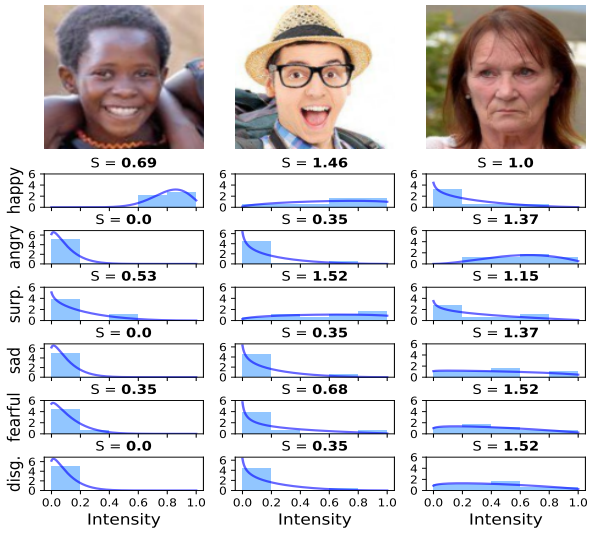

Figure 9. **Annotator agreement is image-specific.** Sample images with varying levels of annotator agreement (measured with entropy S; see Eq. 1, Sec. 3.3 and Fig. 8). The density of the raw ratings are shown with the beta fit for each expression dimension. (Sec. 4.2)

4.2. Annotator behavior and consistency

We begin our analysis of the experimental results by examining annotator behavior. The annotation process was reasonably quick. As expected annotators took more time to simultaneously annotate more expression dimensions (median time per image 25 seconds for the primary (6), 56 seconds for compound (15), 62 seconds for the set of both (21)). Annotators were also less willing to rate images along more dimensions. As seen in figure 7, more annotators gave up after annotating a few images for the compound and 21 expressions tasks, which we take as an indication that such tasks are more demanding.

Next, we considered how often and how intensely each expression is labeled in the dataset. Fig. 6 shows a histogram of how often each intensity rating was used across all images and annotators during the primary expressions annotation task. We then explored how many face images were perceived to exude multiple expressions by calculating the total number of expression dimensions where the median intensity rating meets (equal or larger) a threshold $\tau = 0.5$ (i.e., a majority of annotators indicate that an expression can be perceived at least moderately).

On images rated solely along the six primary expressions we find that 28.7% of images do not satisfy this condition for any expression, 58.5% meets the threshold for one expression, 10.1% for two expressions, 2.3% for three, 0.4% for four, and none with more than four expressions having a median intensity rating ≥ 0.5 for a given face. Thus, we find that single primary expressions are prevalent, and that compound expressions are mostly the superposition of *two* primary expressions and not more.

	Ratings used to Fit Beta Distribution						
	8	7	6	5	4	3	2
angry	1.69	1.78	1.93	2.21	2.81	4.42	10.24
disgusted	2.03	2.18	2.44	2.92	3.84	5.88	11.78
fearful	1.79	1.92	2.14	2.52	3.29	5.11	10.85
happy	1.62	1.71	1.85	2.12	2.73	4.46	10.60
sad	1.80	1.93	2.12	2.48	3.22	4.97	10.83
surprised	1.79	1.89	2.06	2.40	3.16	5.18	12.24

Table 3. **How many annotators are needed for reproducible measurements?** Beta distributions are fit to the data using all possible combinations of c = 2, 3,.., L-1 annotator ratings (Sec. 4.2). Cross-entropy, H_β (Eq. 2, Sec. 3.3), was computed from the left-out annotations. The mean cross-entropy values across all images are shown by expression and the number of annotations used to fit each distribution. Cross-entropy saturates around 5-6 ratings, indicating six annotators are sufficient to collect reproducible ratings.

Our annotation approach lets us examine not only the prevalence of expressions, but also the variability amongst annotations. We find that some expressions yield near-perfect annotator agreement while others receive intensity annotations across the entire scale, indicating that the expression is ambiguous. We modeled this by calculating the entropy S defined in Eq. 1, for each intensity distribution along each expression dimension. Fig. 9 illustrates three sample images with relatively low, medium, and high entropy scores. The histogram of entropy scores across all images annotated for the primary expressions is shown in Fig. 8. Annotators agreed most in their ratings of 'disgusted' (mean=0.82, std=0.37) and agreed least on 'surprised' (mean=0.99, std=0.28).

Since perceived expressions can be subjective and ambiguous, it can be hypothesized that multiple annotators are needed to generate reproducible expression distributions for a face image. To explore this, we conducted an analysis to determine how many annotators are needed to fit expression distributions that effectively model human perception. We calculated the cross-entropy, H_β using Eq. 2, between the beta fit and the left out annotator intensity ratings when using all combinations of c = 2, 3, ..., L-1 ratings where L=9. Table 3 shows the mean H_β values across all combinations of annotator ratings per expression dimension for each image. Entropy values generally decrease as more ratings are used to fit the beta, converging around 5 or 6 ratings for most expressions. The "happy" and "angry" dimensions have comparatively low entropy values whereas "disgusted" has comparatively large entropy values, even when using 8 intensity ratings to fit the beta. Some expression dimensions, where more perception variance exists, may need more annotators when crowdsourcing to generate a stable distribution.

4.3. Benchmarking algorithms

Multidimensional modulated annotator ratings can help us better model the human perception of expression and assess the performance of automated systems. Many modern expression perception algorithms output confidence scores for the 6 primary expressions. Cross entropy (H_d) and absolute distance (M_d) described in Sec. 3.4 can be used to directly compare algorithmic scores with human perception. We applied our benchmark dataset and method to the evaluation of 4 state-of-the-art expression detection algorithms, including two commercial algorithms ('C1', 'C2') and 2 open source academic algorithms (RMN [48], DAN [53]). Each algorithm takes a face image as input and produces an expression vector as output that included at least the 6 primary expressions of interest. We directly compare the algorithmic output with the mean, μ, of the beta distribution. We evaluate the performance of the algorithms using H_d and M_d between μ and the algorithm's predicted value. Our cross-entropy metric specifically accounts for the spread of annotator ratings across the 5 bins.

We also generate annotator ground truth for an additional expression dimension ($d = D + 1$), "neutral". With the rationale that "neutral" could be interpreted as the opposite of any of the non-neutral facial expressions, we formulate a "neutral" raw ratings, per annotator l, per image i, denoted as $v_{D+1} = 4 - (\max([v_1, v_2, ..., v_D]))$, then compute the normalized intensity rating r_{D+1}.

Benchmark results can be found in Fig. 10, where DAN shows the best performance at predicting intensity for primary expressions in terms of absolute distance and cross-entropy metric, followed by C1. We show in Fig. 2 a sample image and the primary expression predictions of the 4 selected algorithms. To compare our metrics with traditional metrics used to benchmark expression detection algorithms, we calculated classification accuracy and F-score using the original one-hot ExpW labels on the 1,000 subset. Results for each algorithm are shown in Table 4 where DAN appear to be the best across half the expressions and metrics. Thus, our evaluation method gives a sharper endorsement to the DAN algorithm.

4.4. Testing the compound expression hypothesis

Are six primary dimensions sufficient to completely characterize a facial expression [17] (Sec. 2)? More stringently: are many/most/all expressions well represented as the compounds of one-two primary expressions [13, 42]? Thanks to our 1,000 annotated face images, where we collected perceptions limiting annotators to, respectively, the 6 primary, the corresponding 15 two-way compound, and the 21=6+15 expressions dimensions, we are now in a position to start exploring these questions.

To this end, we constructed three algorithms to *infer compound expressions from primary ones*. The first is a linear model taking into account *all* primary expressions. The second is a set of 15 linear models, each one of which combines the two normative compound expressions (Table 1). The third is a set of 15 linear models where we selected the two most informative primary expressions for each compound expression. See Fig. 4 for an example.

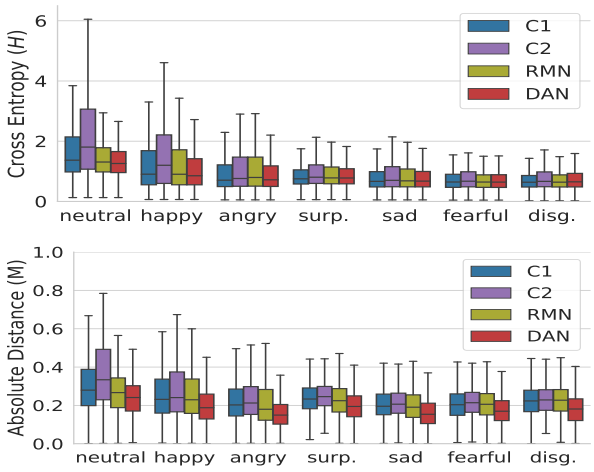

Figure 10. **Expression prediction benchmark.** Four state-of-the-art automated expression prediction algorithms were benchmarked using two metrics: cross entropy (H_d) and absolute distance (M_d) as defined in Sec. 3.4. The box plots show the median (horizontal line), 66% confidence intervals (colored boxes) and 95% confidence intervals (whiskers) (see Sec. 4.3).

	C1		**C2**		**RMN**		**DAN**	
	F_1	*Acc.*	F_1	*Acc.*	F_1	*Acc.*	F_1	*Acc.*
neutral	**0.61**	0.75	0.60	**0.85**	0.48	0.40	0.49	0.38
happy	0.79	0.80	0.78	**0.86**	0.77	0.76	**0.80**	0.80
angry	0.41	0.51	**0.43**	0.46	0.24	0.64	0.31	**0.68**
surprised	**0.49**	0.38	0.41	0.32	0.38	0.40	0.43	**0.46**
sad	0.38	0.28	0.35	0.28	0.42	0.32	**0.49**	**0.40**
fearful	-	-	-	-	-	-	-	-
disgusted	0.00	0.00	0.00	0.00	0.07	0.05	**0.15**	**0.14**

Table 4. **Comparing traditional benchmark metrics**. We compare the results of 4 algorithms on original ExpW ground truth [60] using standard multi-class classification metrics: accuracy and F-score (Sec. 4.3). The highest score per metric per expression dimension is highlighted. Both Accuracy and F-score are to be maximized.[1]

We conducted a 10-fold cross-validation experiment where we fit the models using 90% of images and predicted the μ (see Sec 3.3). Each time, we measured the Mean Absolute Error (MAE) between the prediction and the actual value (we express MAE as percent of the $[0, 1]$ dynamic range). We found MAE=6.2% for the full model, and MAE=7.5% and 7.8% respectively for the two models utilizing only two dimensions. These findings suggest that compound expressions may be predicted accurately using the six primary dimensions, with a small amount of information lost in the process. Further analysis is needed to better characterize this small effect.

5. Conclusion and discussion

We proposed a novel method to collect and model multi-dimensional modulated facial expression annotations. The method improves upon previous work because it does not rely on expert annotators, multiple expressions are simultaneously annotated (up to 21) with their perceived level of intensity, and expression ambiguity can be measured. Using our method, we annotated a diverse set of 1,000 in-the-wild face images. We were then able to benchmark two commercial and two academic expression prediction algorithms using two different metrics and found the DAN [53] algorithm to be the best overall. Our metrics refine our understanding of algorithm performance by considering the underlying distribution of human expression perception.

Additional findings emerge from our study. First, 5-6 crowdsourced annotators are sufficient for achieving reproducible measurements. Second, while the expression of most in-the-wild face images is well characterized by one of Ekman's six primary dimensions, some faces require two dimensions to be characterized. Third, annotator perceptions are best measured using a modulated, rather than binary, scale. Fourth, annotators will agree on the perception of most faces; however, for a number of expressions perception is ambiguous, with annotations spreading over many intensities. Any method that benchmarks expression prediction algorithms has to take such ambiguities into account. Fifth, while we find that reducing expression annotation to six primary dimensions is quite effective, we observe that a small amount of information is lost in the process – understanding how this happens will require further investigation.

Future directions of research in this area include further exploring algorithmic techniques that allow models to learn from both clear and ambiguous facial expressions. Another interesting question is exploring individual annotator behavior, e.g. understanding whether annotators of different age, gender, and ethnicity may perceive facial expressions differently in a systematic way and whether some annotators may have richer perception than others.

Ethics discussion. Measuring human perception of facial expression from images helps build better benchmarks for automating the perception of facial expression in machines. This, in turn, will enable engineers to build machines that can better interact with human users, thus enabling machines to address a wider range of human needs. Alongside the obvious benefits, this technology presents risks, including change in the patterns of human social interaction, control of privacy, the potential for racial bias, and potentially unforeseen economic impact due to rapid technological change [23, 27].

Acknowledgements: We are grateful to Aleix Martinez, Umit Keles, Ralph Adolphs, Stefano Soatto, Ayanna Howard, and Wei Li for guidance in reading the relevant cross-disciplinary literature and providing insightful comments on the manuscript.

[1]The inclusion criteria for image selection (Sec. 4.1) yielded too few samples of "Fear", an expression underrepresented in the original EXPW labels (≈1% of dataset). Yet, the distribution of "Fear" annotations we obtained closely resembles that of other expressions (see Fig. 6), thus highlighting the information gain of using multidimensional modulated annotations.

References

[1] D. Acharya, Z. Huang, D. Pani Paudel, and L. Van Gool. Covariance pooling for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 367–374, 2018. 2, 4

[2] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest*, 20(1):1–68, 2019. 2

[3] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283, 2016. 3

[4] M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Measuring facial expressions by computer image analysis. *Psychophysiology*, 36(2):253–263, 1999. 1, 2, 3

[5] C. Blank, S. Zaman, A. Wesley, P. Tsiamyrtzis, D. R. Da Cunha Silva, R. Gutierrez-Osuna, G. Mark, and I. Pavlidis. *Emotional Footprints of Email Interruptions*, page 1–12. Association for Computing Machinery, New York, NY, USA, 2020. 4

[6] D. Bryant and A. Howard. A comparative analysis of emotion-detecting ai systems with respect to algorithm performance and dataset diversity. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 377–382, New York, NY, USA, 2019. Association for Computing Machinery. 4

[7] M. I. Conley, D. V. Dellarco, E. Rubien-Thomas, A. O. Cohen, A. Cervera, N. Tottenham, and B. Casey. The racially diverse affective expression (radiate) face stimulus set. *Psychiatry research*, 270:1059–1067, 2018. 3

[8] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001. 1

[9] C. Darwin. *The expression of the emotions in man and animals*. University of Chicago press, 2015. 1

[10] A. Dawel, L. Wright, J. Irons, R. Dumbleton, R. Palermo, R. O'Kearney, and E. McKone. Perceived emotion genuineness: normative ratings for popular facial expression stimuli and the development of perceived-as-genuine and perceived-as-fake sets. *Behavior research methods*, 49(4):1539–1562, 2017. 3

[11] S. Deng, Y. Xiong, M. Wang, W. Xia, and S. Soatto. Harnessing unrecognizable faces for improving face recognition. *arXiv preprint arXiv:2106.04112*, 2021. 6

[12] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Classifying facial actions. *IEEE Transactions on pattern analysis and machine intelligence*, 21(10):974–989, 1999. 1, 2

[13] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014. 2, 4, 8

[14] H. L. Egger, D. S. Pine, E. Nelson, E. Leibenluft, M. Ernst, K. E. Towbin, and A. Angold. The nimh child emotional faces picture set (nimh-chefs): a new set of children's facial emotion stimuli. *International journal of methods in psychiatric research*, 20(3):145–156, 2011. 3

[15] P. Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992. 2

[16] P. Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993. 1

[17] P. Ekman and E. L. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997. 2, 3, 8

[18] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. 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. 3

[19] A. H. Farzaneh and X. Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2402–2411, 2021. 2, 4

[20] E. Goeleven, R. De Raedt, L. Leyman, and B. Verschuere. The karolinska directed emotional faces: a validation study. *Cognition and emotion*, 22(6):1094–1118, 2008. 3

[21] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pages 117–124. Springer, 2013. 3

[22] J. Guo, Z. Lei, J. Wan, E. Avots, N. Hajarolasvadi, B. Knyazev, A. Kuharenko, J. C. S. J. Junior, X. Baró, H. Demirel, et al. Dominant and complementary emotion recognition from still images of faces. *IEEE Access*, 6:26391–26403, 2018. 3

[23] Y. N. Harari. *21 Lessons for the 21st Century*. Random House, 2018. 8

[24] N. Hirschberg, L. E. Jones, and M. Haggerty. What's in a face: Individual differences in face perception. *Journal of Research in Personality*, 12(4):488–499, 1978. 3

[25] H. Hoffmann, H. Kessler, T. Eppel, S. Rukavina, and H. C. Traue. Expression intensity, gender and facial emotion recognition: Women recognize only subtle facial emotions better than men. *Acta psychologica*, 135(3):278–283, 2010. 2

[26] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions, volume 2*, volume 289. John wiley & sons, 1995. 4

[27] S. Jonze. Her (film). *Annapurna Pictures*, 2013. 8

[28] T. Kanade et al. *Computer recognition of human faces*, volume 47. Birkhäuser Basel, 1977. 2

[29] U. Keles, C. Lin, and R. Adolphs. A cautionary note on predicting social judgments from faces with deep neural networks, Jan 2021. 1

[30] U. Keles, C. Lin, and R. Adolphs. A cautionary note on predicting social judgments from faces with deep neural networks. *Affective Science*, pages 1–17, 2021. 2

[31] D. Keltner, D. Sauter, J. Tracy, and A. Cowen. Emotional expression: Advances in basic emotion theory. *Journal of nonverbal behavior*, pages 1–28, 2019. 2, 3

[32] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388, 2010. 3

[33] J. S. Lerner and D. Keltner. Fear, anger, and risk. *Journal of personality and social psychology*, 81(1):146, 2001. 2

[34] H. Li, N. Wang, X. Ding, X. Yang, and X. Gao. Adaptively learning facial expression representation via cf labels and distillation. *IEEE Transactions on Image Processing*, 30:2016–2028, 2021. 2

[35] S. Li and W. Deng. Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. *International Journal of Computer Vision*, 127(6):884–906, 2019. 3, 4

[36] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 3

[37] V. LoBue and C. Thrasher. The child affective facial expression (cafe) set: Validity and reliability from untrained adults. *Frontiers in psychology*, 5:1532, 2015. 3

[38] J. Lu, X. Xie, and R. Zhang. Focusing on appraisals: How and why anger and fear influence driving risk perception. *Journal of safety research*, 45:65–73, 2013. 2

[39] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pages 200–205. IEEE, 1998. 3

[40] D. S. Ma, J. Correll, and B. Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, 47(4):1122–1135, 2015. 3

[41] K. Mase. Recognition of facial expression from optical flow. *IEICE TRANSACTIONS on Information and Systems*, 74(10):3474–3483, 1991. 1, 2

[42] S. McCloud. *Making comics: Storytelling secrets of comics, manga and graphic novels*. Harper New York, 2006. 2, 4, 8

[43] A. S. Meuwissen, J. E. Anderson, and P. D. Zelazo. The creation and validation of the developmental emotional faces stimulus set. *Behavior research methods*, 49(3):960–966, 2017. 3

[44] 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*, 10(1):18–31, 2017. 3

[45] B. Montagne, R. P. Kessels, E. H. De Haan, and D. I. Perrett. The emotion recognition task: A paradigm to measure the perception of facial emotional expressions at different intensities. *Perceptual and motor skills*, 104(2):589–598, 2007. 2

[46] I. J. Myung. Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, 47(1):90–100, 2003. 5

[47] C. Padgett and G. W. Cottrell. Representing face images for emotion classification. *Advances in neural information processing systems*, pages 894–900, 1997. 1, 2

[48] L. Pham, T. H. Vu, and T. A. Tran. Facial expression recognition using residual masking network. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4513–4519. IEEE, 2021. 7

[49] R. Plutchik. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984:197–219, 1984. 2, 4

[50] A. Todorov, A. N. Mandisodza, A. Goren, and C. C. Hall. Inferences of competence from faces predict election outcomes. *Science*, 308(5728):1623–1626, 2005. 2

[51] N. Tottenham, J. W. Tanaka, A. C. Leon, T. McCarry, M. Nurse, T. A. Hare, D. J. Marcus, A. Westerlund, B. Casey, and C. Nelson. The nimstim set of facial expressions: judgments from untrained research participants. *Psychiatry research*, 168(3):242–249, 2009. 3

[52] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6897–6906, 2020. 2, 4

[53] Z. Wen, W. Lin, T. Wang, and G. Xu. Distract your attention: Multi-head cross attention network for facial expression recognition. *arXiv preprint arXiv:2109.07270*, 2021. 7, 8

[54] J. Willis and A. Todorov. First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological science*, 17(7):592–598, 2006. 2

[55] F. Xue, Q. Wang, and G. Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3601–3610, 2021. 2, 4

[56] Y. Yacoob and L. S. Davis. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on pattern analysis and machine intelligence*, 18(6):636–642, 1996. 1, 2

[57] S.-T. Yeh et al. Using trapezoidal rule for the area under a curve calculation. *Proceedings of the 27th Annual SAS® User Group International (SUGI'02)*, pages 1–5, 2002. 5

[58] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006. 3

[59] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2008. 1

[60] Z. Zhang, P. Luo, C.-C. Loy, and X. Tang. Learning social relation traits from face images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3631–3639, 2015. 2, 3, 5, 8